

---

# **MASTER THESIS**

---

Mr.  
**Sebastian Bittrich**

**Toward a Unified Assessment  
Protocol for Local and Global  
Protein Structure Stability and  
Quality Using Knowledge-Based  
Potentials**

2014



# **MASTER THESIS**

---

## **Toward a Unified Assessment Protocol for Local and Global Protein Structure Stability and Quality Using Knowledge-Based Potentials**

Author:

**Sebastian Bittrich**

Course of studies:

Molecular Biology/Bioinformatics

Seminar group:

MO12w1-M

First examiner:

Prof. Dr. rer. nat. Dirk Labudde

Second examiner:

M.Sc. Florian Heinke

Mittweida, August 2014

*"What's the use of a candle that casts no light?"*

*"It is a lesson", Armen said, "the last lesson we must learn before we don our maester's chains. The glass candle is meant to represent truth and learning, rare and beautiful and fragile things. It is made in the shape of a candle to remind us that a maester must cast light wherever he serves, and it is sharp to remind us that knowledge can be dangerous. Wise men may grow arrogant in their wisdom, but a maester must always remain humble. The glass candle reminds us of that as well. Even after he has said his vow and donned his chain and gone forth to serve, a maester will think back on the darkness of his vigil and remember how nothing that he did could make the candle burn... for even with knowledge, some things are not possible."*

---

*A Feast for Crows, Prologue*  
by George R. R. Martin

---

## Bibliographic description

Bittrich, Sebastian: Toward a Unified Assessment Protocol for Local and Global Protein Structure Stability and Quality Using Knowledge-Based Potentials, 69 pages, 12 figures, Hochschule Mittweida (FH), Department of Mathematics, Natural and Computer sciences

Master Thesis, 2014

## Abstract

Proteins are involved in almost every aspect of life, mediating a wide range of cellular tasks. The protein sequence dictates the spatial arrangement of the residues and thus ultimately the function of a protein. Huge effort is put into cumbersome structure elucidation experiments which obtain models describing the observed spatial conformation of a protein, enabling users to predict their function, to understand their mode of action or to design tailored drugs to cure disease caused by misfolded or misregulated proteins.

However, the result of structure determination experiments are merely models of reality, made under simplifying assumptions - sometimes containing major undetected errors. On the other hand, such experiments are resource demanding and they cannot supply the actual demand. Thus, scientists are predicting the structure of proteins *in silico*, resulting in models that are even more prone to error.

In consequence, the structure biologists search after a practicable definition of structure quality and over the last two decades several model quality assessment programs emerged, measuring the local and global quality of peculiar structures. Seven representatives were studied, regarding the paradigms they follow and the features they use to describe the quality of residues. Their predications were compared, showing that there is almost no common ground among the tools. Is there a way to combine their statements anyway?

Finally, the accumulated knowledge was used to design a novel evaluation tool, addressing problems previously spotted. Thereby, high quality of its predication as well as superior usability was key. The strategy was compared to existing approaches and evaluated on suitable datasets.

---

## Zusammenfassung

Fast jede Funktion einer Zelle wird durch Proteine vermittelt. Ihre einzigartige Abfolge von Aminosäuren diktiert das räumliche Arrangement ihrer Bausteine und resultiert letztendlich in definierten Funktionen, die das Protein übernimmt. Ein Gros an Ressourcen wird für Strukturaufklärungsexperimente angewendet, in denen man die räumliche Anordnung der Reste eines Proteins zu bestimmen versucht. Solche Modelle geben Auskunft über die Funktion eines Proteins, seine Wirkungsweise und ermöglichen das maßgeschneiderte Entwickeln von Medikamenten.

Man darf jedoch nicht vergessen, dass die Ergebnisse von Strukturaufklärungsexperimenten nur Modelle der Realität darstellen, die unter vereinfachenden Annahmen gemacht wurden und sogar manchmal komplett falsch sind. Weiterhin sind solche Experimente sehr zeit- und kostenaufwendig und können nicht den wachsenden Bedarf decken. Deshalb gibt es Strukturvorhersagedienste, die *in silico* versuchen, Modelle zu berechnen, welche aber noch fehleranfälliger sind.

Wissenschaftler suchen nach einem anwendbaren Maß für die Strukturgüte und in den letzten 25 Jahren wurden zahlreiche Proteinstrukturevaluationsprogramme entwickelt, die die lokale und globale Qualität von Strukturen zu bewerten versuchen. Sieben Vertreter wurden genauer untersucht hinsichtlich ihrer Ansätze und den Kategorien, die sie nutzen, um die Qualität zu quantifizieren. Nach dem Vergleich ihrer Ausgaben zeigte sich, dass ihre Bewertungen überraschenderweise sehr gegenläufig sind. Gibt es dennoch einen Weg, eine gemeinsame Aussage aus mehreren Tools zu bestimmen?

Zu guter Letzt wurde das gewonnene Wissen in einen eigenen Strukturrevaluationsdienst kanalisiert, der Fehler etablierter Tools vermeidet. Ziel sollte sowohl eine treffende Bewertung der Strukturqualität sein als auch eine hohe Benutzerfreundlichkeit. Abschließend wurde das Vorgehen mit den etablierten Ansätzen verglichen und mittels geeigneter Datensätze evaluiert.

---

## **Acknowledgments**

Gratitude is owed to M.Sc. Florian Heinke who sacrificially supervised this work, offered help while knowing his way around seemingly all aspects of bioinformatics and motivated me during each meeting anew. Prof. Dr. rer. nat. Dirk Labudde gave me the opportunity to study this field and I started enjoying it a long time ago; also he provided valuable feedback and the opportunity to exchange ideas and opinions with other students writing their master thesis. Furthermore, I am truly grateful to Annemarie Bittrich, Rick Schubert, Silvio Oswald and Rico Hellwig who proofread this text. Last but not least, I would like to thank my family, friends and fellow students for encouraging and supporting me. You have led me to write these lines at this very moment.





# I. Contents

<b>Contents</b>	<b>I</b>
<b>List of Figures and Tables</b>	<b>II</b>
<b>1 Motivation</b>	<b>1</b>
<b>2 Theoretical Principles</b>	<b>3</b>
2.1 Protein Structure Determination . . . . .	4
2.2 Structural Quality Assurance . . . . .	5
2.3 Protein Structure Prediction . . . . .	7
2.4 Critical Assessment of Protein Structure Prediction . . . . .	9
<b>3 Model Quality Assessment Programs</b>	<b>11</b>
3.1 Verify3D . . . . .	11
3.2 PROCHECK . . . . .	12
3.3 VADAR . . . . .	13
3.4 PROSESS . . . . .	13
3.5 Mean Force Potentials in General . . . . .	14
3.6 ProSA . . . . .	15
3.7 ANOLEA . . . . .	15
3.8 QMEAN . . . . .	16
3.9 Feature Comparison . . . . .	17
3.10 Progress in the Field of MQAP . . . . .	19
<b>4 Protein Energy Profiling</b>	<b>21</b>
4.1 Mathematical Principles . . . . .	21
4.2 Energy Profile Suite . . . . .	23
<b>5 Materials and Methods</b>	<b>25</b>
5.1 Creating Datasets . . . . .	25
5.2 Extension of BioJava . . . . .	26
5.3 Training of the Evaluation Model . . . . .	28
5.4 Deployment as Web Service . . . . .	29
5.5 Calculation of Consensus Energy Profiles . . . . .	31
<b>6 Results</b>	<b>33</b>
6.1 Correlations between CASP10 Data and Local Residue Descriptors . . . . .	33
6.2 Comparison of the MQAP . . . . .	35
6.3 Performance on CASP Datasets . . . . .	41
6.4 Approximating Energy Profiles Using the Consensus Method . . . . .	45
<b>7 Discussion</b>	<b>47</b>
<b>8 Outlook</b>	<b>51</b>

<b>9 Summary</b>	<b>53</b>
<b>Bibliography</b>	<b>55</b>

## II. List of Figures and Tables

### Figures

2.1 Protein folding and energy [Kim et al., 2013] . . . . .	3
2.2 Schema of x-ray diffraction and R-factor calculation [Laskowski, 2005] . . . . .	6
2.3 Flowchart of homology modeling - adapted from [Liu et al., 2011] . . . . .	8
4.1 Intra-molecular occurring interactions - adapted from [Heinke and Labudde, 2012] . .	22
5.1 Superposition of an experimentally derived structure and one corresponding model . .	26
5.2 Sample output for 1IXB . . . . .	30
6.1 Structure visualization for Calmodulin (1EXR) . . . . .	37
6.2 Local evaluation for Calmodulin (1EXR) . . . . .	38
6.3 Structure visualization for Xylose Isomerase (1MUW) . . . . .	39
6.4 Local evaluation for Xylose Isomerase (1MUW) . . . . .	40
6.5 Variety of evaluations for CASP10 target T0645 . . . . .	43
6.6 Variety of evaluations for CASP9 target T0515 . . . . .	44

### Tables

3.1 MQAP global properties . . . . .	17
3.2 MQAP local assessment features . . . . .	18
6.1 Spearman's rank correlation coefficient for the EP-based approach on the CASP10 dataset . . . . .	33
6.2 Spearman's rank correlation coefficient for QMEAN . . . . .	34
6.3 MQAP local evaluation correlation matrix . . . . .	35
6.4 Spearman's rank correlation coefficient for the EP-based approach on the CASP9 dataset . . . . .	41
6.5 Spearman's rank correlation coefficient for QMEAN on the CASP9 dataset . . . . .	42
6.6 Descriptive statistics for methods predicting the native structure's energy profile . . . .	46



# 1 Motivation

Almost every cellular process and every aspect of life revolves around proteins. 20,000 to 25,000 proteins are found in the human organism [Kim et al., 2013]. With this omnipresence, problems arise though: absent, nonfunctional or misregulated proteins results in a variety of diseases ranging from cancer and autism [Zhao et al., 2014] over diabetes insipidus [Heinke and Labudde, 2012] to Alzheimer's disease [Kang et al., 1987]. In consequence, to understand such illnesses and to potentially design drugs, we need particularized knowledge of the properties of a protein [Kuntz, 1992].

In 1894, Fischer formulated the connection between the spatial conformation of a protein and its function. Analogous to a lock and key, an enzyme provides a geometric shape complementary to molecules processed by it [Fischer, 1894]. Roughly 40 years later, Wu [Wu, 1931] as well as Mirksy and Pauling [Mirsky and Pauling, 1936] showed independently that this ordered structure can be corrupted by manipulating the chemical environment of a protein. This denaturation results in conformational changes of the protein and a consequent loss of its initial function. Furthermore, stepwise restoration of the physiological condition results in folding of the protein once more. For one century, structure was equated with function. However, with the emergence of ways to determine the exact structure of a protein [Kendrew et al., 1958] examples of proteins concurrently emerged which featured significant disorder and were partially kind of denatured; and yet this odd properties facilitate the function of these proteins. They are no static objects and their inherent flexibility allows them e.g. to bind to a wide range of cellular targets, governing i.a. signal transduction or regulatory processes. Furthermore, eukaryotic cells tend to feature way more proteins exhibiting disorder than prokaryotic ones, implying their importance in evolution and in setting prokaryotes and eukaryotes apart. Deletion of disordered proteins is also more likely to be lethal than knocking out non-disordered ones [Dunker and Obradovic, 2001; Dunker and Kriwacki, 2011].

A protein structure describes the arrangement of its components and enables dedicated scientists to further investigate peculiar proteins. Therefore, elaborate structure determination experiments are done, yielding molecular photographs which try to describe the observation of the experiment as good as possible. Yet they are merely snapshots embracing only a small part of a possibly bigger picture. It is obvious that such models cannot capture the spatial rearrangements occurring within a protein structure. It demands not only enormous expertise to derive good, useful structures, but no matter how accurate they may be, they will never span all the processes potentially taking place. It is not surprising that the structure of one individual protein is determined multiple times: by independent teams, under varying conditions, in presence of other binding partners or simply because better technology is available. Even if we cannot capture all aspects of a protein in one run, in a way by combining multiple single images we can still converge towards reality. Another way to give this fact consideration are energy

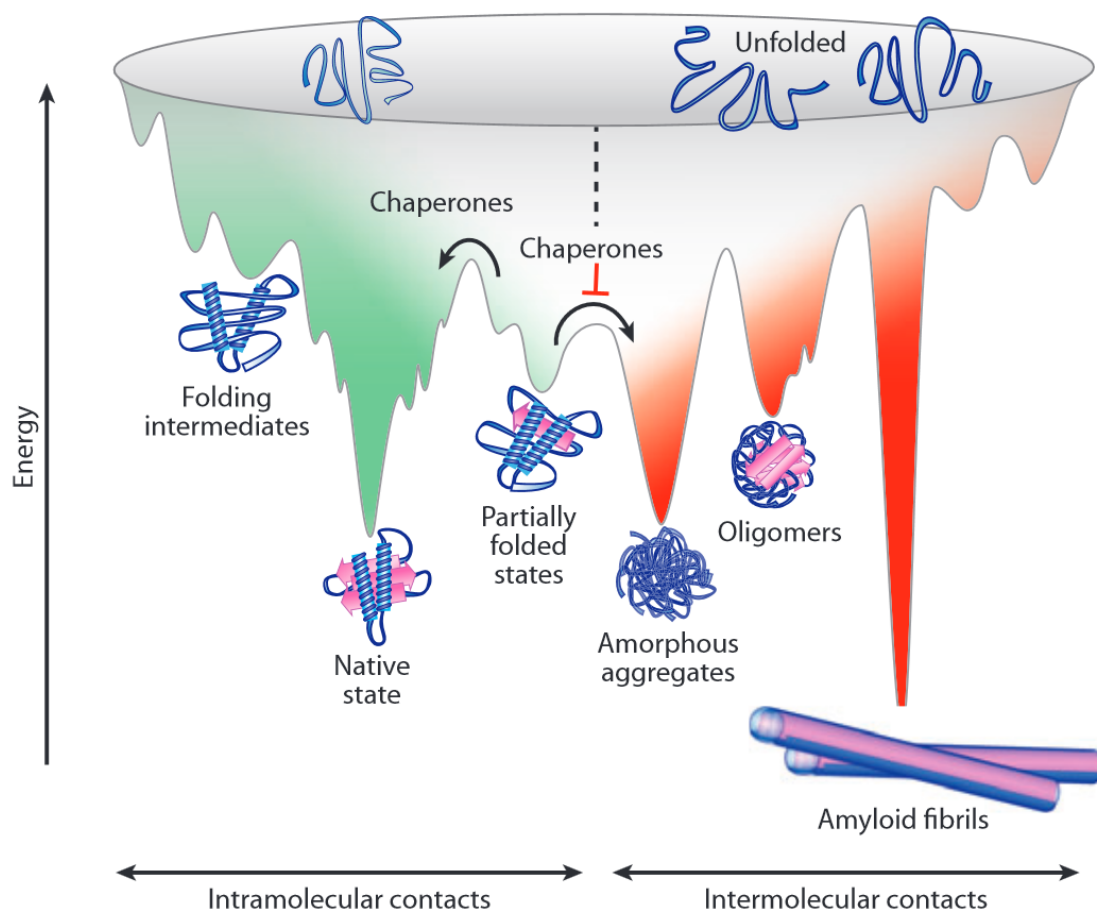
models like molecular dynamics or coarse-grained knowledge-based potentials which try to describe the protein's energy landscape and the rearrangements the protein may undergo.

The quality of derived models limits their uses and for demanding applications like spotting binding sites or studying the molecular mechanisms, structures of high quality are mandatory. But how can the quality of a protein structure be quantified?

This work can roughly be partitioned into two aspects. First, ways to measure the quality of the protein models were studied - using information obtained directly by the structure determination experiment or by the use of dedicated protein structure evaluation tools. On the other hand, a novel quality assessment program was designed using the knowledge obtained previously and it was tested on several datasets regarding the quality of its predication and to enable a comparison with the other tools.

Throughout the text the term *amino acid* refers to one of the 20 standard amino acids and its particular features. In contrast to that, *residue* is related to single instances in a particular protein. Names of mentioned Java-classes or interfaces are printed *italic*. Due to default of a catchy name, the here discussed evaluation approach is referred to as *EP-based*.

## 2 Theoretical Principles



**Figure 2.1:** Protein folding and energy [Kim et al., 2013]

Newly synthesized proteins strive after favorable states of minimal energy. Chaperones help other proteins to fold correctly in the first place, preserve the native conformation of preexisting ones and prevent unfavorable aggregation of (partially) misfolded proteins (shaded red).

Scientists use two major aspects to describe proteins: the protein sequence is the succession of amino acids synthesized during translation; the protein structure however captures the native, three-dimensional conformation a protein exhibits after folding. Depending on the arrangement of the residues, the protein's energy changes (Figure 2.1). It is proposed that the energy landscape is funnel-shaped and a protein in a native or near-native state is located in a local minimum regarding their energy value [Anfinsen, 1973; Dill and MacCallum, 2012; Kim et al., 2013; Nelson and Cox, 2010]. Chaperones are a class of proteins aiding the correct folding of other proteins or helping to overcome energetic barriers of partially folded proteins stuck in a local minimum [Kim et al., 2013]. Interestingly, there are also proteins which lack any ordered structure; they are flexible and only exhibit an ordered spatial conformation when they are interacting with

other macromolecules. Almost all disordered proteins are involved in signal transduction processes and need to bind to a huge set of different molecules. Theories such as the *protein trinity* [Dunker and Obradovic, 2001] or the *protein-quartet* [Uversky, 2002] propose that this intrinsic disorder governs the possibility to bind to several ligands and that the function of a protein is not only the result of its ordered spatial structure, but is also realized by partially or wholly disordered conformations and, furthermore, any transitions between these states [Dunker and Obradovic, 2001; Kovacevic, 2012]. The native state is not strictly the one of minimal energy, but rather a compromise between an advantageous thermodynamical state and conformational flexibility needed to function properly [Kim et al., 2013].

Several databases collect information brought up by the biologists. Protein structures are e.g. publicly accessible via the protein data bank (**PDB**) [Bernstein et al., 1977] and can be visualized by dedicated tools such as PyMOL (<http://www.pymol.org/>).

## 2.1 Protein Structure Determination

While the sequence is quiet easily to obtain - with next generation sequencing even accelerating the process - tertiary structure determination however is a resource-consuming procedure without guaranteed success. For each protein an x-ray diffraction or NMR experiment is needed, itself demanding huge expertise from the operator. Proteins are purified and experimental data are obtained which are finally transformed into a spatial model of the protein, explaining the observations as good as possible.

X-ray diffraction utilizes crystals of purified proteins which causes single molecules to arrange themselves in a repeating array. When this crystal is stricken by a beam of x-ray radiation, the beam is scattered into several single instances (Figure 2.2). Thereby, the real space is imaged on the reciprocal space and each atom results in a reflection, although more than one atom can reflect the beam to the exact same position, resulting in varying intensity though. Reflections are characterized by their amplitude and phase, but the phase is unknown and can not be directly derived from the experimental data [Kendrew et al., 1958; Nelson and Cox, 2010; Rhodes, 2006; Wlodawer et al., 2007]. To determine the phases, e.g. libraries of high-resolution data of small molecules can be consulted or a scaffold can be created by methods such as homology modeling and improved afterwards with knowledge of the experimental data [Giorgetti et al., 2005; Raimondo et al., 2006; Wlodawer et al., 2007]. So the diffraction pattern is used to generate a coarse electron density map which is used in an iterative refinement process for protein structure model building, trying to arrange all the residues in a way to optimally explain the diffraction pattern [Kendrew et al., 1958; Nelson and Cox, 2010; Rhodes, 2006; Wlodawer et al., 2007]. Even though the majority of structures are solved by x-ray diffraction the success rates are rather low: 5% of sequences targeted for structure determination by x-ray are in fact successfully published, as only 20% get to the



crystallization trials in the first place. Membrane proteins are even harder to express, purify and crystallize, as they are not dissolved in water but in a hydrophobic lipid bilayer [Thornton, 2001].

Since proteins are seldom embedded in a crystal under physiologic conditions [Doye and Poon, 2006] and since it is often quite challenging to generate useful crystals in the first place, other methodologies such as **NMR** (nuclear magnetic resonance) were developed which study proteins in solution or solid state and furthermore allow observations over small timescales. Nuclei with an odd number of protons or neutrons have a nonzero nuclear spin. Due to application of a magnetic field, the spins can be aligned and nuclei possess two distinguishable energetic states now: one favorable and one disadvantageous of higher energy opposing the external field. A suitable radio frequency pulse can force the transition between the two states. The phenomenon of nuclear magnetic resonance can be exploited to obtain information about the environment in which certain nuclei are located in - they are slightly shifted away from the expected values. These chemical shift data is finally used for model building. Usually, an ensemble of roughly 20 models is obtained by NMR which are included in one PDB file or are condensed into one model by averaging and by subsequent energy minimization [Laskowski, 2005; Nelson and Cox, 2010; Wüthrich, 1990].

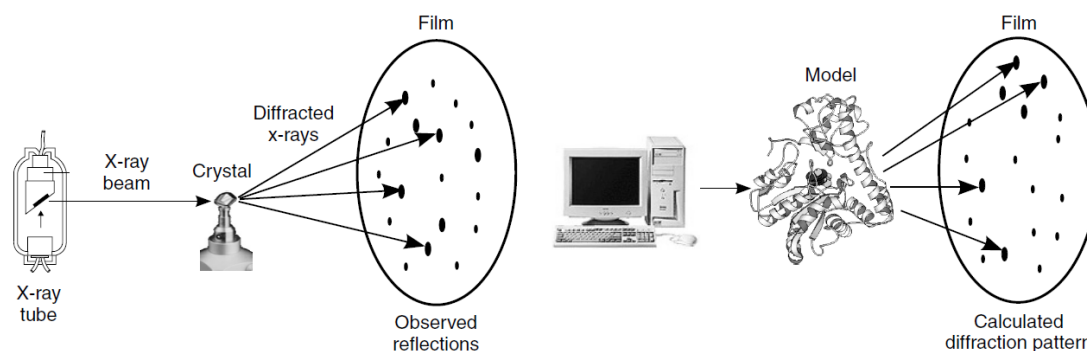
But one should keep in mind that proteins are dynamic systems being in motion, reacting to changes in their chemical environment and possibly alter their conformation upon interacting with other proteins, DNA or other ligands. Their behavior cannot necessarily be captured by a snapshot, just as a stage play cannot be captured all-embracingly by a photograph. Each protein structure is solely a model of reality - likely containing defects [Laskowski, 2005].

## 2.2 Structural Quality Assurance

Though scientist try to publish models as good as possible in the first place, constantly new, better protein structures emerge, replacing old ones. Now and then completely wrong protein models [Read et al., 2011] or even knowingly false contributions are revealed [Borrell, 2009]. It is quite difficult to assess the quality of certain models, to compare them and to spot the most reliable. As structure quality dictates possible applications [Kryshtafovych et al., 2011; Kuntz, 1992; Liu et al., 2011], it is crucial to assess the quality as accurately as possible and to get to know weaknesses and limitations. Thus, since 2011 the PDB provides quality reports for each new entry available for authors and users. They aim at being easy to understand even for non-experts while providing much more detailed further information for interested users [Read et al., 2011]. Probably the most basic measure for the global quality of a protein structure is the resolution. It can be considered a value expressing how much experimental data was incorporated into the model, determining the minimum distance of structural components which can still be distinguished in the electron density map. The resolution in measured

in Å and a certain value indicates that reflections caused by equivalent, parallel crystal planes located so far apart are included in the model. Unfortunately, such reflections are scattered at wider angles, decrease in intensity and are generally harder to measure, but they still provide valuable extra information. However, there is no unambiguous definition: some scientists refer to the biggest scattering angle yielding a complete data set, others use the widest single diffraction angle [Laskowski, 2005; Rhodes, 2006; Wlodawer et al., 2007].

The R-factor measures how well the calculated model is able to explain the observed experimental data. Therefore, a refraction pattern based on the model is computed and the degree of disagreement between the two pattern is expressed as R-factor (Figure 2.2). Models with a R-factor below 0.2 are considered as being finished in their refinement and being ready to be published as reliable structure. Random or entirely false models give R-factors of 0.4 to 0.6. However, even completely wrong models can exhibit convincing R-factors, e.g. when far too many water molecules are included [Laskowski, 2005]. Roughly one water molecule per residue is reasonable and they should only be placed when they result in plausible hydrogen bonds [Brändén and Alwyn Jones, 1990]. Due to the manipulability of the R-factor, the free R-factor or  $R_{\text{free}}$  was designed



**Figure 2.2:** Schema of x-ray diffraction and R-factor calculation [Laskowski, 2005]

When a protein crystal is stricken by an x-ray beam, the characteristic diffraction pattern can be used to derive the structure of the protein. The R-factor captures the agreement between the diffraction pattern observed in the experiment and the predicted one based on the derived model.

as more meaningful measure of the agreement between experimental data and the derived model. The methodology is almost the same as for the standard R-factor, but a small part of the protein of around 5–10% is excluded from the model refinement process and solely used for the calculation of the  $R_{\text{free}}$  value, thus detaching refinement and measuring the quality of fit, preventing any bias introduced by the model building process directly influencing the  $R_{\text{free}}$  calculation [Brünger, 1992; Laskowski, 2005; Read et al., 2011]. The  $R_{\text{free}}$  is usually larger than the R-factor and even harder to interpret, but values above 0.4 indicate possible problems [Brünger, 1997].

While previous variables capture the global quality of proteins, the B-factor indicates protein dynamics on the level of individual atoms. It measures how smeared out the electron density is, respectively how precise the coordinate of a particular atom is. Thermal

motion of the atoms or disorder in this part of the protein cause this disadvantageous movement [Laskowski, 2005; Rhodes, 2006].

$$B = 8\pi^2 \langle U_{eq}^2 \rangle \quad (2.1)$$

Thereby,  $\langle U_{eq}^2 \rangle$  describes the mean squared displacement of an atom. It assumes isotropic vibration in all directions. More sophisticated experimental methods like cryo-preparation and the use of synchrotron radiation make it is also possible to obtain by far more precise anisotropic displacement parameters, covering all directions individually. For their visualization oftentimes ellipsoids are used [Merritt, 1999]. The higher the B-factor is, the higher is the positional uncertainty of that atom. One should avoid relying on such atoms and also exclude all atoms whose B-factors exceed  $40.0 \text{ \AA}^2$  [Laskowski, 2005; Rhodes, 2006].

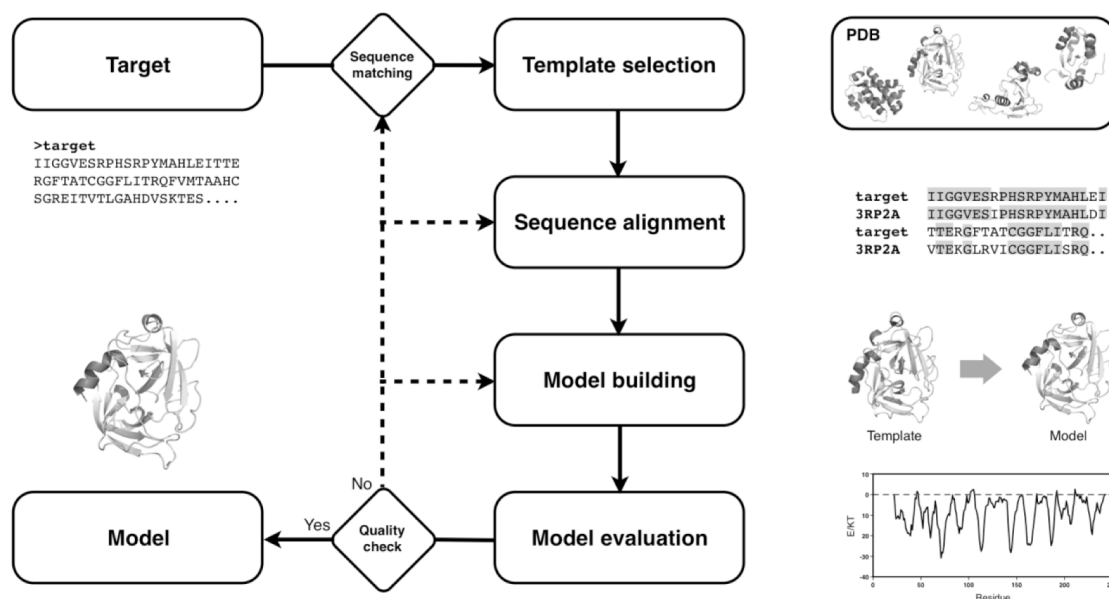
Presented quality measures are directly connected to the x-ray diffraction method; as NMR-experiments lack scatter angles or diffraction patterns, they also lack any conventional inherent global quality measure. Thus, their quality needs to be assessed otherwise.

In summary, even structures with good quality measures can still have misinterpreted the experimental data. Possibly, an atom is placed wrong based on the electron density map and the assumptions made, but the B-factor could state high precision for its position though. A reasonable indicator for reliable protein structures are however multiple structures deposited in the PDB, which were obtained from independent teams and under varying conditions. The structures are likely correct, when all these experiments derived a similar model from the observed data [Laskowski, 2005].

## 2.3 Protein Structure Prediction

Whereas the number of known protein sequences still increases rapidly, it is still cumbersome to determine the structure of a protein. E.g., if a scientist wants to do mutation studies on a structure, he would not only have to synthesize the desired protein structure but also to vanquish the difficulties during structure determination. It is reasonable to model a so far unknown structure with knowledge of physicochemical principles (called *ab initio*) or by constructing models with knowledge of the structure of closely related proteins (*de novo* modeling) [Liu et al., 2011].

Homology modeling (Figure 2.3) is an approach of the latter category and utilizes the fact that homologous proteins share a mutual ancestor. They feature high sequence identity and therefore their structure should be similar as well. The PDB (or any database containing known structures) is used to find homologous proteins by FASTA [Pearson, 1990], BLAST [Altschul et al., 1990], PSI-BLAST [Altschul, 1997] or approaches involving hidden Markov models [Soding, 2004]. Depending on the envisioned use of the model, a suitable template is selected [Liu et al., 2011] - not necessarily the one exhibiting highest sequence identity, some methods even use multiple templates, some



**Figure 2.3:** Flowchart of homology modeling - adapted from [Liu et al., 2011]

A database is scanned for suitable a template with high sequence identity to the target sequence. The best alignment between query sequence and the selected template is computed; their known structure is used to derive fragments leading to the model. By a suitable method, the quality is assessed and refined until a satisfying degree of quality is achieved or no further improvement can be observed.

authors suggest to pay close attention to the milieu of the structure like pH and present ions or ligands [Sadowski and Jones, 2007]. Both sequences are aligned by standard algorithms such as Needleman-Wunsch [Needleman and Wunsch, 1970] or Smith-Waterman [Smith and Waterman, 1981]. However, especially when dealing with low sequence identities, multiple sequence alignments or methods incorporating structural information from homologous proteins can perform far better. After template selection and alignment, small peptides are used as building blocks to compute the backbone of the protein. Loops are constructed with the aid of a database containing the structure of peptide fragments. Side chain coordinates can be predicted by their intrinsic preference or observations in the template structure. Restraints and constraints can guide the model building process as well, by minimizing the number of violations. In particular, loops are hard to model due to commonly occurring insertions and deletions. Backbone-dependent rotamer libraries can aid the correct placement of side chains. In an iterative process, badly modeled regions can be refined by molecular dynamics or simplified force fields [Liu et al., 2011]. A wide range of tools exist to evaluate the final model involving environment classes [Eisenberg et al., 1997], geometry and stereochemistry [Berjanskii et al., 2010; Laskowski et al., 1993; Willard, 2003] or potentials of mean force [Benkert et al., 2011; Melo et al., 1997; Wiederstein and Sippl, 2007]. However, the performance of homology modeling is directly limited by sequence similarity to the template structure and the quality of the sequence alignment [Kihara et al., 2009]. Sequence similarities below 40% require a manual inspection of the alignment.

The selection of unsuitable templates or misalignment can have devastating results. Finally, a related structure is not known for every sequence [Liu et al., 2011].

## 2.4 Critical Assessment of Protein Structure Prediction

The critical assessment of protein structure prediction is a biennial experiment (**CASP**) evaluating the quality of existing protein structure modeling approaches. A set of protein sequences of soon to be published structures is provided. Participating teams try to predict the three-dimensional structure of these sequences, while their real structure is being derived simultaneously by experimental methods. With the deadline, the one experimentally determined structure is released and about 500 theoretical models from the attendees can be compared. Targets vary in difficulty, just as prediction methods vary in how well they perform on certain targets. In the most objective way, the currently best performing methods are spotted, which outperformed others and worked especially well among the whole set of CASP targets. Over the course of subsequent CASP runs, the development of the existing methodology can be monitored and the whole community gets to know which approaches perform well overall or on certain problematic targets, resulting in the possibility to adopt knowledge from other teams. Rather than mere competition, CASP aims at exchanging ideas and accelerating the evolution of protein structure modeling tools [Cozzetto et al., 2007; Kryshtafovych and Fidelis, 2009; Kryshtafovych et al., 2011, 2014b; Moult et al., 1995]. Over the course of the years, CASP started also including other fields: most notably is the model quality assessment category, where participating teams are encouraged not only to submit models but also the state how reliable the model can be considered on local and global level. In summary, the procedure for the modeling category is adopted for the model quality assessment branch: ranking the different approaches as well as monitoring the progress over the years [Cozzetto et al., 2007; Kryshtafovych and Fidelis, 2009; Kryshtafovych et al., 2011, 2014a].

To assess the quality of each theoretical model, it is compared to the experimental derived structure published in the PDB. Two structures can be aligned and their degree of similarity can be expressed in numerical values. Commonly used is the root-mean-square deviation (**RMSD**) which describes the average distance between pairs of backbone atoms.

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N \delta_i^2} \quad (2.2)$$

Thereby is  $\delta_i$  the distance between one of the  $N$  equivalent atom pairs. Most alignment algorithms aim at translating and rotating on structure onto the other in a manner to minimize the RMSD [Kabsch, 1976, 1978; Krissinel and Henrick, 2004; Maiti et al., 2004]. However, the RMSD approach struggles heavily as soon as proteins with multiple domains are considered. Particularly due to the CASP experiment more sophisticated

global alignment scores were developed. The global distance test (**GDT**) states how well two structures match each other in a manner similar to a percentage. Nowadays it is one of the standard evaluation measures of CASP [Kryshtafovych et al., 2014b; Sadowski and Jones, 2007; Zemla, 2003].

$$GDT = \frac{1}{4} \sum_{t=1,2,4,8} \frac{f(t) \cdot 100}{N} \quad (2.3)$$

Therefore, a set of thresholds  $t$  is defined, usually 1, 2, 4 and 8 Å. The fraction of atom pairs  $f(t)$  falling below that certain threshold is then normalized by the total number of atom pairs in the alignment. In context of the CASP experiment, models with values above 80% can be considered successfully modeled [Forrest et al., 2006; Giorgetti et al., 2005; Read and Chavali, 2007]. Such GDT values indicate a RMSD around 1 Å to the experimental derived structure, rendering the model suitable for demanding problems such as spotting ligand binding sites, studying molecular mechanisms causing functionality or disease, solving structures by molecular replacement and, ultimately, even drug discovery. Mediocre quality models (GDT above 50, RMSD to the experimental structure around 2–3 Å) allow the user to find probable active sites, virtual screening and predicting the impact of mutations closely related to illnesses. Last but not least, models of low quality are still useful for rough domain annotation and sophisticated guesses on the molecular function of the protein [Kryshtafovych et al., 2011].

CASP targets are sometimes not directly comparable, one may be easy to model because the structure of a close homologue is known, whereas the other lacks any suitable template. Z-scores can be used to compare two populations of data with different range of values and scaling by normalizing the data. The expected value  $\mu$  and the standard deviation (**SD**)  $\sigma$  of each set of values is used to standardize each data point  $x$ .

$$z = \frac{x - \mu}{\sigma} \quad (2.4)$$

As a result, it states how many standard deviations a value is off it the average of the data.

### 3 Model Quality Assessment Programs

However, quality values originating directly from experimental methods are not as significant as they may seem at first glance: there is no consistent definition for terms such as resolution, the interpretation of some values strongly depends on others and it is even possible to sugarcoat variables like the R-factor. Correct interpretation of all available information demands huge expertise [Brändén and Alwyn Jones, 1990; Kryshtafovych and Fidelis, 2009; Laskowski, 2005]. Furthermore, x-ray model building involves minimization of a function strongly related to the R-factor which is also used to state the final accuracy of the model, thus resulting in unfavorable circularity which should optimally be avoided [Eisenberg et al., 1997]. There is a significant demand for model quality assessment programs (**MQAP**) estimating the reliability of a structure in a unbiased, easy interpretable, fast and meaningful manner.

Two major motivations exist to develop MQAP: some aim at helping crystallographers approaching the problem from the experimental point of view while others originated from theoretical methods, developed to predict protein structures *in silico*. The latter mostly follows the paradigm that a native structure exhibits a nearly minimal total energy. Furthermore, *in silico* modeled structures do not provide any quality indication at all, making it even harder for modelers and users to estimate their reliability. Structure prediction pipelines like SWISS-MODEL [Arnold et al., 2006; Biasini et al., 2014] roughly sculpture a structure and iteratively perform refinement cycles hoping to improve the accuracy of the prediction. Thereby, various MQAPs like QMEAN [Benkert et al., 2009a, 2011, 2008, 2009b], ANOLEA [Melo et al., 1997; Melo and Feytmans, 1997, 1998] and GROMOS [Oostenbrink et al., 2004] are consulted, indicating whether certain structural modifications in the newest refinement cycles caused a desirable change in the overall quality of the model or how reliable the final model is. This implies that a superior MQAP can enhance current modeling strategies - not by directly performing predictions but by serving as meta-predictor selecting native confirmations.

#### 3.1 Verify3D

The 1991 established approach Verify3D originated from assumptions made concerning the inverse folding problem known as finding suitable protein sequences for a given structure [Bowie et al., 1991].

For each residue three values are assessed: the fraction of buried surface area of the side chain, the fraction of polar groups or water adjacent to the side chain and the local secondary structure. These values are then used to assign each residue an environment class, transforming the complex 3D structure into an easy to handle, one-dimensional string of categories similar to a protein sequence. For example, there is

a differentiation between buried, partially buried and exposed side chains concerning their accessible surface area (**ASA**) as well as a further subdivision with respect to the polarity of the environment. Every one of the 20 standard amino acids also has a certain preference whether it occurs in a helix, sheet or coil region. Residues confronted with uncommon environment get assigned low 3D-1D scores. Finally, the most favorable combination of protein sequence and newly formed environment string is sought by a dynamic programming alignment, called 3D compatibility search. Gaps are allowed and even locally unfavorable combinations are tolerated as long as they are overcome by high scoring regions [Bowie et al., 1991; Eisenberg et al., 1997].

The alignment score can finally be used to calculate a Z-score, expressing how many SD is the observed alignment score above that of other sequences featuring a similar length [Gribskov et al., 1990]. Values above 7 nearly always indicate the same general fold as the structure represented by the profile. As a result, the authors claim to fuse two distinct lines of protein science: sequence comparison and conformational energy calculation using stereochemical and packing features of each residue. The method can distinguish between correct and globally misfolded structures and is able to spot locally erroneous regions [Bowie et al., 1991; Eisenberg et al., 1997]. It is also notable that Verify3D is available for download whereas some MQAP are only accessible as Web service, rendering them quite inconvenient when it comes to processing large amounts of data.

## 3.2 PROCHECK

PROCHECK was published in 1992 and is considered a well-established, easy to use MQAP, thus being utilized for protein structure assessment even nowadays. PROCHECK aims at helping crystallographers finding common mistakes like mislabeled atoms or incorrect L/D stereochemical labels which are automatically correct when calling the program. Several PostScript plots summarize the results [Laskowski et al., 1993, 1996].

Similar to Verify3D certain preferences are determined and then used to check whether the structure meets these expectations. While Verify3D uses the environment preference of each amino acid, PROCHECK operates at atomic level assessing bond lengths and angles as well as atom contacts. A set of small molecules [Engh and Huber, 1991] was used to determine expected values and SD of bond lengths and angles occurring for all heavy atoms of the protein backbone. When evaluating a structure, the difference between observation and expected value is computed and normalized by the SD. Residues whose features aberrate more than 0.5 SD are labeled. Pairs of  $\phi$ - and  $\psi$ -angles can be assessed by the Ramachandran-Ramakrishnan-Sasisekharan-plot [Ramachandran et al., 1963] - dihedral angles located in unfavorable or disallowed regions of the plots indicate errors. Also identified are nonbonded interactions, being defined as the closest contact between two residues which are less than 4 Å apart but feature a sequence separation of more than four bonds [Laskowski et al., 1993, 1996].

The G-factor states how normal or close to the expected value a certain property is,



when it is being compared to structures with a similar resolution. E.g. the Ramachandran-plot was divided into 45 x 45 cells aiming at deriving the probability of a certain amino acid exhibiting that particular combination of  $\phi$ - and  $\psi$ -angles. G-factors are log-odds derived from the distributions of stereochemical properties, whereby low values indicate residues showing unfavorable behavior like disallowed dihedral angles. Because of the nature of log-odds, it is possible to formulate meaningful averages of the residue-wise values yielding a global predication [Laskowski et al., 1993, 1996]. A standalone version can be downloaded, furthermore it is accessible via the PDBsum (<http://www.ebi.ac.uk/pdbsum/>) [Laskowski, 2001].

### 3.3 VADAR

Both methods - Verify3D and PROCHECK - were combined by **VADAR** (volume, area, dihedral angle reporter) in 2003. In total, 15 different tools were merged trying to meld the best properties of all approaches [Willard, 2003].

Considered aspects include backbone and side chain torsion angles, observable secondary structure as well as propensity, ASA, hydrogen bond energies, solvation energy, stereochemical features and Verify3D's profiling approach. Quite unique is the inspection of excluded volume as calculated by the Voronoi polyhedra method of Richards [Richards, 1977]. It represents the volume occupied by the amino acid's atomic radii as well as its nearest neighbors. The fractional volume normally takes values close to 1.0. However, values exceeding 1.2 indicate interior cavities, whereas such below 0.8 usually occur in compressed or poorly refined parts of the structure. All of the structural parameters get finally combined to one score ranging from 0 to 9, whereby values dropping below 5 mark problematic parts of the structure [Willard, 2003].

VADAR does not provide any form of composite score for global quality assessment. However, the average of most contemplated features is provided, but it is up to the user to compare them to annotated expected values [Chiche et al., 1990; Miller et al., 1987; Morris et al., 1992; Richards, 1977] and derive a meaningful predication.

### 3.4 PROSESS

Seven years later, **PROSESS** (protein structure evaluation suite & server) got back to VADAR's reasoning of combining different existing tools in a useful manner - existing approaches were evaluated concerning their usefulness and whether they are in keeping with the times. Over 100 measurement criteria were incorporated, some directly derived from Verify3D, PROCHECK or VADAR [Berjanskii et al., 2010].

Notable is the support for raw experimental data such as **NOE**-based (nuclear Overhauser effect) distance restraints and NMR chemical shifts whose consistency with the submitted structure can be tested. Furthermore, GeNMR [Berjanskii et al., 2009] implements interaction and solvation energy terms and SuperPose [Maiti et al., 2004]

searches for homologous structures and uses their evaluation to gain further information. Usually, a run takes 3–5 minutes and presents the global and local assessment by extensive use of graphs, charts, tables and color-coded as well as color-mapped structure images [Berjanskii et al., 2010].

PROSESS formulates six major quality categories: overall, covalent and geometric, non-covalent/packing, torsion angle, chemical shift and NOE quality. Each term is compared to a high-quality reference dataset by means of Z-scores, which are then scaled to range from 0 to 10. Instead of simple averaging, a rather sophisticated weighting and calibration scheme is applied to make the scores more sensitive when distinguishing structures of consimilar quality [Berjanskii et al., 2010].

### 3.5 Mean Force Potentials in General

While the previous methods derive statistical preferences directly, it is also conceivable to wrap the probability of a certain feature into an energetic term according to the inverse Boltzmann's law yielding so called potentials of mean force [Hendlich et al., 1990; Sippl, 1990, 1993a, 1995].

$$E(f) = -k_B T \cdot \ln \frac{P_{obs}(f)}{P_{ref}(f)} \quad (3.1)$$

Thereby is  $P_{obs}(f)$  the probability of the feature  $f$  derived from the set of structures, whereas  $P_{ref}(f)$  indicates the probability of the reference state.  $k_B$  denotes the Boltzmann constant and  $T$  the temperature [Li, 2013; Sippl, 1990]. There are energy models or dedicated MQAP calculating contact energies between interacting residues or solvation energies in dependency of residue's accessibility. Interpretation as contact energy is obvious since interactions within the protein also directly govern initial folding [Anfinsen, 1973]. Their individual realization differs, but all follow the hypothesis of native proteins presenting minimal total energies. On residue level high potentials indicate unfavorable states such as hydrophobic groups being exposed to the solvent [Benkert et al., 2008; Heinke et al., 2013; Sippl, 1993a]. Even though the calculated values do not resemble the actual physical values, their usefulness is proven [Jones et al., 1992; Sippl, 1993a,b]. Knowledge-based mean force potentials (**MFP**) are proven in use when it comes to threading and fold recognition [Jones et al., 1992; Sippl, 1993a], homology modeling [Sippl, 1993b], molecular docking [Verkhivker et al., 1995] and last but not least protein structure evaluation.

However, there are many ways to parameterize an energy model. What terms should be considered? What dataset is used to derive the statistics? Which distance thresholds decide whether two residues interact according to the model? Is a certain sequential separation between residues demanded - e.g. residues have to be at least four indexes in the protein sequence apart in order to be able to interact in the first place? Thus, even though most energy models are derived from Sippl's assumptions uttered in 1990 [Hendlich et al., 1990; Sippl, 1990], there is extreme variability in the exact implementation.

## 3.6 ProSA

In contrast to previously mentioned MQAP, **ProSA** (protein structure analysis) follows a completely different paradigm, postulating that local atom clashes or other trivial steric principles are not as important as the correct arrangement of all residues in a holistic, global manner when it comes to the protein's 3D structure - local flaws can be compensated by a correct tertiary structure in general [Domingues et al., 2000; Sippl, 1993b, 1995; Wiederstein and Sippl, 2007]. Each residue is represented by a energy value describing its environment and how well it matches the amino acid's preferences.

ProSA uses primarily  $C_\alpha$  atoms to represent residues which is useful when dealing with minimal structure models missing side chains or even  $C_\beta$  atoms. However, in the standalone-version also statistics for the  $C_\beta$  representation exist which is suspected to carry even more information. Interaction distances of 4 to 15 Å are considered, eliminating both very close and too spacious contacts. The profile is finally smoothed over a window size of 39 residues. The authors also mentioned that their approach was designed for globular structures and though membrane proteins show similar behavior, their predication is little understood [Domingues et al., 2000; Sippl, 1993b, 1995; Wiederstein and Sippl, 2007].

To assess the global quality of a structure, again a Z-score is utilized, whereby the observed energies are compared to an energy distribution obtained from random conformations [Sippl, 1993b, 1995]. Flawed models are pointed out by Z-scores away from typical values. Since these depend on the protein size, a plot showing Z-scores of all structures in the PDB is provided in order of comparability [Wiederstein and Sippl, 2007]. ProSA is available for download and as Web service.

## 3.7 ANOLEA

While ProSA chooses one atom for each residue to represent the group as good as possible, **ANOLEA** (atomic non-local environment assessment) established in 1997 deals with each heavy atom individually. This should allow a more fine-grained view, especially since it eliminates the need to find a suitable representation for a residue, an as it turns out avoidable procedure losing a decent amount of information by simplifying too much [Melo et al., 1997; Melo and Feytmans, 1997, 1998].

Unfortunately, by the more differentiated approach the number of observations for certain heavy atom types shrink, rendering the statistics unreliable. Therefore, 40 different groups of atoms sharing physico-chemical properties were established. These depend on their connectivity, chemical nature and whether they are side chain atoms or located in the backbone. In this energy model, two residues interact in a non-local manner when they are located closer than 7 Å and the residues are farther away than 11 residues in the protein sequence or are part of different chains. The final energy of an amino acid is the sum of the energy of all its atoms, averaged over a window of 5 residues [Melo et al., 1997; Melo and Feytmans, 1997, 1998].

The model's total energy is then again used to form a Z-score based on the dataset used for training. It shows a moderate correlation to the resolution of the structure. Very high Z-scores occur for structures exhibiting an incorrect stereochemistry. Furthermore, upon submitting a job to the Web server, a threshold for high-energy residues can be specified. Energy values above the threshold are marked in the output and the fraction of unfavorable high-energy residues is given as secondary global quality predication [Melo et al., 1997; Melo and Feytmans, 1997, 1998].

### 3.8 QMEAN

In 2008, **QMEAN** (qualitative model energy analysis) was released which is a linear combination of potentials, agreement terms, relative ASA and the ratio of unassigned secondary structure using a window size of nine residues [Benkert et al., 2009a, 2011, 2008, 2009b].

Interaction, solvation and torsion energy terms are incorporated. Contact energies are realized using an amino acid- and secondary structure-specific representation by  $C_{\beta}$  atoms. Contacting residues need to be at least 4 residues apart concerning the protein sequence and a distance range of 3–15 Å with a bin size of 0.5 Å is used. The solvation energy is similarly implemented using a sphere of 9 Å and processing the fraction of found residues in this sphere divided by the maximum number observed in the dataset used for training. Furthermore, the novel torsion energy calculated for 3 consecutive amino acids helped further enhancing the quality of this approach. The agreement values compare the ASA and secondary structure assigned by DSSP [Joosten et al., 2011; Kabsch and Sander, 1983] and predictions of ACCpro [Pollastri et al., 2002] and PSIPRED [McGuffin et al., 2000] solely based on the sequence - postulating contradictions occur in unreliable parts of the protein. Among the here presented programs, only QMEAN performs well on structures determined by both experiment [Benkert et al., 2008] and *in silico* modeling [Kryshtafovych et al., 2011]. Unmatched is the quality of the visualization of the results as well as the usability of the server - supporting archives containing sets of structures and a convenient option to assess the quality of multiple chains. Major drawback (like when dealing with PROSESS) is a high computation times of several minutes whereas other service are capable of almost instant response [Benkert et al., 2009a, 2011, 2008, 2009b].

Last but not least, QMEAN also uses Z-scores to derive a statement on the global quality of a model. Both the contributing terms and the final QMEAN-score are compared to high-resolution X-ray proteins of comparable size ( $\pm 10\%$  the length of the query protein's sequence). Remarkable is that two reference datasets are used, differentiating between single chains and oligomeric assemblies. The higher the QMEAN Z-score, the more reliable is the model [Benkert et al., 2011].

### 3.9 Feature Comparison

Most MQAP use Z-scores for global assessment (Table 3.1) whereby some raw scores is compared to scores obtained during training. Small proteins tend to exhibit lower overall quality just because less stabilizing interactions can occur. Thus, global assessment scores and even Z-scores depend on the size of a structure and Z-scores are often calculated for proteins of comparable size only. PROCHECK uses log-odds but still features a global quality score. PROSESS provides six categorical scores but lacks any composite values summing up the results. Only VADAR does not provide any form of global quality indicator.

**Table 3.1:** MQAP global properties

MQAP	year	paradigm	global score
Verify3D	1991	compatibility between structure and sequence	Z-score
PROCHECK	1992	preference of stereochemical features	log-odds
VADAR	2003	enhance existing tools by ASA and energy terms	-
PROSESS	2010	reevaluate and combine most useful existing tools	scaled categorical scores
ProSA	1990	knowledge-based potential	Z-score
ANOLEA	1997	knowledge-based potential	Z-score
QMEAN	2008	knowledge-based potential	composite score, Z-score

On local level however the tools differ significantly more (Table 3.2). To define Verify3D's environment classes, the ASA, the observable secondary structure and the occurrence of polar groups in direct neighborhood of each residue is esteemed. Contrary, PROCHECK consults stereochemical properties like bond lengths and torsion angles as well as a check for unfavorable non-bound atoms e.g. those who are located closer than the sum of their Van der Waals radii. Verify3D and PROCHECK are complementary concerning the contemplated values making it reasonable for VADAR as well as PROSESS to combine both approaches. Furthermore, they add energetic values for observable hydrogen bonds and solvent interactions to their approach. Actually, PROSESS also looks at residue-residue interaction energies, thus covering all of the afore mentioned categories. ProSA and ANOLEA are solely energy models and do not refer

to any other properties for their evaluation. However, their energy model is quite sophisticated and well-documented, whereas that of VADAR and PROSESS is anything but transparent as they provide no calculation rules or actual usable values in the output files. This inaccessibility of energetic values contrasts the aim to reasonably cover all categories during PROSESS's design. QMEAN takes a mainly energetic point of view, though also stating that trivial properties like the ASA and the fraction of loop around a residue are important variables [Benkert et al., 2008]. The here presented methods are

**Table 3.2:** MQAP local assessment features

feature	Verify3D	PROCHECK	VADAR	PROSESS	ProSA	ANOLEA	QMEAN
bond lengths & angles		✓	✓	✓			
torsion angles		✓	✓	✓			✓
non-bound atom contacts		✓	✓	✓			
polar groups in environment	✓		✓	✓			
hydrogen-bond energy			✓	✓			
solvation energy			✓	✓	✓		✓
interaction energy				✓	✓	✓	✓
accessible surface area	✓		✓	✓			✓
secondary structure	✓		✓	✓			✓

only a small fraction of the available tools. While the most common paradigms are covered, way more programs were developed. The quality of the afore mentioned MQAPs was however stated on multiple independent sources, they are frequently used and relatively easy accessible by automatable scripts. Worth mentioning is also the existence of dedicated evaluation tools for DNA as well as RNA [Gendron et al., 2001] or the HETZE program for hetero groups binding to proteins [Kleywegt and Jones, 1998]. On the other hand, there are tools performing exceptionally well in the CASP environment, but they are useless for everyday quality assessment tasks. Some are consensus methods, combining knowledge from one entire CASP target, following the paradigm that common observations are likely correct. However, under normal circumstances one does not have access to 500 models from a wide range of extremely versatile methodologies, but instead much smaller sets at best [Kryshtafovych et al., 2014a]. So it is questionable how useful such methods are even when they perform extraordinary well on this particular data.

### 3.10 Progress in the Field of MQAP

MQAP should be able to rank multiple models according to their global quality. And in fact many existing tools are capable of arranging sets of models according to their global reliability. But while it is feasible to directly compare the quality of multiple models with each other, it is still hard to assess the global quality of only one model and to predict a quality measure like the GDT individually, which then enables correct ranking of models when ranked among other individually predicted values [Cozzetto et al., 2007]. Without knowledge of other available models, global quality is still challenging even though decent improvements were made recently [Kryshtafovych and Fidelis, 2009; Kryshtafovych et al., 2011].

In contrast to that, the significance of quality assessment on residue level decrease over past CASP-runs. However, that is not because contemplated tools worsen, but they did not improve either while targets steadily increase in complexity. It is assumed that slight modifications of existing tools will not result in any significant improvement. Some revolutionary idea is necessary to significantly promote development [Kryshtafovych et al., 2011]. Again, non-clustering methods struggle in comparison to consensus approaches combining knowledge from multiple models [Kryshtafovych et al., 2014a].

Especially approaches dealing with single models are at a disadvantage but they are the most important as they are virtually applicable for every model quality assessment problem [Kryshtafovych and Fidelis, 2009; Kryshtafovych et al., 2014a]. Meta-approaches aim at combining knowledge derived from existing methodologies and merging their results into a single composite score e.g. by machine-learning techniques [Kryshtafovych and Fidelis, 2009] or linear combination [Sadowski and Jones, 2007].





## 4 Protein Energy Profiling

To approach the process of protein folding, the thermodynamic hypothesis was formulated: proteins are to some degree self-organizing systems striving after minimal free energy which promotes the folding process into the functional, native structure. Scientists tried to derive a Hamiltonian function describing the process as function of intramolecular interactions and the interplay with the solvent. Disagreement remains on how far such a function can be simplified. Quantum chemistry is capable of approaching the problem strictly bottom-up. However, there are still major limits regarding the degree of realizable accuracy [Clementi, 2008; Sippl, 1993a]. In contrast, statistical physics tend toward the problem top-down, utilizing observable potentials in known protein structures and applying this knowledge to approximate the energy of other protein structures [Sippl, 1993a].

There are various ways to parametrize an energy model. However, none is perfect or suitable for all tasks - they are merely models of reality. Thus, a bunch of energy models was developed and studied, each exhibiting individual strengths and drawbacks. One approach of this category are so-called energy profiles (**EP**) [Dressel, 2008; Heinke et al., 2013; Heinke and Labudde, 2012].

### 4.1 Mathematical Principles

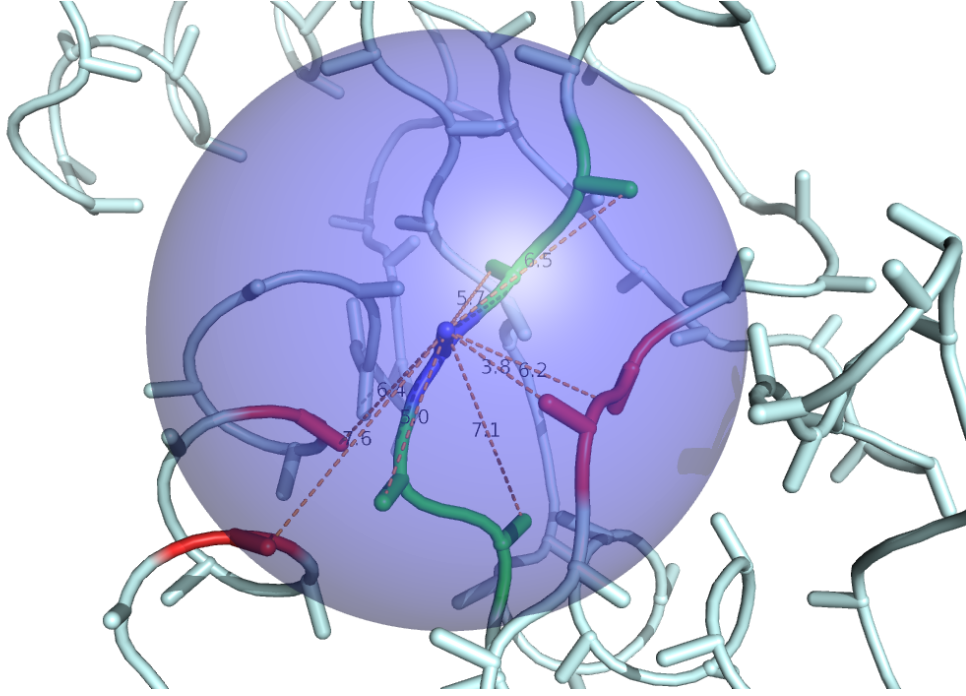
Each amino acid has a certain preference whether it tends to be buried in the hydrophobic core of the protein or rather be exposed to the solvent. In consequence, spacious, hydrophobic amino acids result in energetically unfavorable states when they are forced to interact with the polar solvent. The same goes for highly charged residues confronted with unpolar environments. Differentiation for a residue  $i$  between being buried or located inside and being exposed or outside is realized by

$$f(i) = \begin{cases} n_{aa,buried}++ & \text{if } \|C_{\alpha,i} - c\| < 5 \vee (C_{\alpha,i} - C_{\beta,i})(C_{\alpha,i} - c) < 0 \\ n_{aa,exposed}++ & \text{else} \end{cases} \quad (4.1)$$

with  $c$  being the center of mass of all  $C_{\alpha}$  atoms less than 5 Å away from  $i$ .

For both globular and  $\alpha$ -helical membrane proteins, these preferences for each amino acid were derived and based on the inverse Boltzmann principle. The pseudoenergy  $e_i$  of a residue  $i$  can be approximated by

$$e_i = -k_B T \cdot \ln \frac{n_{aa,buried}}{n_{aa,exposed}} \quad (4.2)$$



**Figure 4.1:** Intra-molecular occurring interactions - adapted from [Heinke and Labudde, 2012]

During calculation each individual residue of the protein is considered (blue). Sequentially close (green) as well as spatially close residues from other parts of the protein (red) are found in the 8 Å sphere and contribute the residue's energy value.

In this context  $k_B$  and  $T$  are constants, thus they can be omitted. The total energy  $E_i$  of a residue  $i$  embedded in a protein structure  $S$  is the sum of all pairwise pseudoenergies

$$E_i = \sum_{j \in S \setminus i} g(i, j)(e_i + e_j) \quad (4.3)$$

with the contact function  $g(i, j)$  being defined as

$$g(i, j) = \begin{cases} 1 & \text{if } \|C_{\beta,i} - C_{\beta,j}\|_{D_E} < 8 \\ 0 & \text{else} \end{cases} \quad (4.4)$$

But residues do not only interact with the solvent, they are also in contact with other residues within the structure. Thus, a potential  $e_{ij}^*$  is defined

$$e_{ij}^* = -\ln \frac{n_{aa_i, aa_j}}{f_{aa_i} f_{aa_j} N} \quad (4.5)$$

with  $n_{aa_i, aa_j}$  being the number of observed contacts between the amino acid of  $i$  and  $j$  in the training dataset. The relative frequency of the amino acids of  $i$  and  $j$  are referred to as  $f_{aa_i}$  and  $f_{aa_j}$ ,  $N$  is the total number of interacting residues. However, it remains unclear how to combine both values exactly. Thus, calculation of the contact energy was implemented but both terms were kept separated, so that in fact two energy profiles can

be computed - one containing solvation energies and a second one dealing with contact energies.

In summary, the complex 3D arrangement of the protein's residues is transformed to an easy-to-handle, yet informative array of energy values [Dressel, 2008; Heinke et al., 2013; Heinke and Labudde, 2012].

## 4.2 Energy Profile Suite

Major advantage in comparison with other energy models is the availability of an entire toolbox called **eProS** (energy profile suite), for dealing with previously generated energy profiles or even predicting them entirely based on the protein sequence as well as a database of pre-calculated profiles for all PDB entries [Heinke et al., 2013].

*eAlign* is a pairwise alignment approach for energy profiles motivated by Needleman-Wunsch [Needleman and Wunsch, 1970], Smith-Waterman [Smith and Waterman, 1981] and the profiling approach which led to Verify3D [Bowie et al., 1991]. The significance of an alignment is covered by the distance score (**dScore**), comparing the observed alignment score to the optimal score as well as the average permutation score derived by rearranging and realigning energy profiles [Heinke et al., 2013; Heinke and Labudde, 2012].

Similar to GOR's secondary structure prediction methodology [Garnier et al., 1996; Gibrat et al., 1987; Kloczkowski et al., 2002], *eGOR* aims at predicting discretized energy profiles based on protein sequences. In an information theory-based approach certain compositions of neighboring residues can be linked to knowledge derived from existing energy profiles [Heinke et al., 2013].



## 5 Materials and Methods

The knowledge of existing MQAP should be combined with the energy profile approach to create a novel protein structure evaluation methodology. It is known that the energy profile approach assigns low energies to residues deeply embedded in the structure, which indicates stability of this particular residue. In contrast, residues of coil regions exposed to the solvent usually exhibit high energy values, implying unfavorable conformations and high positional flexibility.

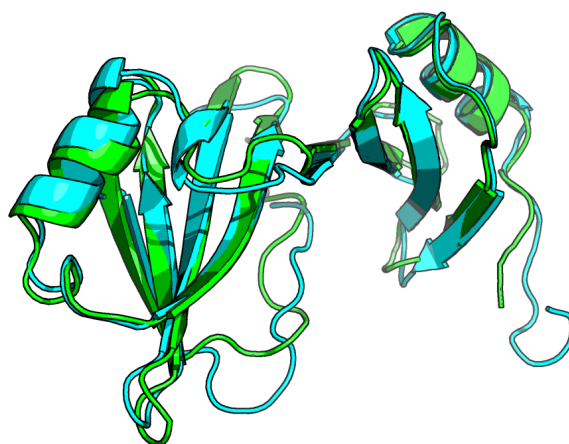
By comparing the predications of the existing tools on suitable datasets, concepts successfully spotting errors could be extracted. In consequence, features of residues (descriptors) were formulated, which served as indicators of the residue's quality and stability. This set of descriptors was finally used to perform a regression, trying to approximate errors actually observed in structure models.

### 5.1 Creating Datasets

For further studies an independent dataset was created, aiming at representing high-quality structures deposited at the PDB while not having any bias towards certain folds or secondary structure elements. PDB-REPRDB [Noguchi, 2001] was used to create a non-redundant dataset of globular structures with a resolution below 1 Å. Proteins missing side chain coordinates, containing nucleic acids or non-standard amino acids were excluded. Clustering took place for sequence identities above 30% or structural similarity below 10 Å. 56 structures were finally selected.

Local installations of PROCHECK and ProSA were utilized while the remaining MQAPs automatically received tasks by querying their web interface. Output files were collected, parsed and their results arranged to match possible offset especially occurring for the very first or last residues. However, 16 structures could not be processed by at least one tool, mostly if they contain multiple chains, implying a quiet low reliability even when dealing with well-formatted obtained directly from the PDB.

In contrast to this dataset containing high-quality structures determined by experimental methods, a second dataset was created representing theoretical models mostly predicted by homology modeling. All structures from CASP10 [Moult et al., 2014] were used to represent models of potentially lower quality. However, the dataset contains more than 40.000 models. Per target five models were selected randomly, though they must have a GDT above 20 to ensure reasonable sequence coverage as well as quality. E.g. there are structures which are wholly unfolded, just being a linear chain of amino acids. Not only superposition with the experimentally derived structures (Figure 5.1) enabled a global quality assessment in form of the GDT, the distance between all corresponding residue pairs can be used as local quality measure. Deviation within the protein core is minimal for good models. However, coil regions facing the solvent and



**Figure 5.1:** Superposition of an experimentally derived structure and one corresponding model

Visualization of the PDB structure (teal) of CASP10 target T0644 aligned by PyMOL with the predicted model T0644TS024\_1 (green). Especially regions exhibiting an ordered secondary structure are located at similar positions in both structures while coil compartments tend to deviate in their coordinates. The distance between equivalent atoms was used to quantify the error of the model at that position.

lacking stabilizing interactions tend to be modeled with less certainty, resulting in greater distances between residue pairs. Distances were limited to 15 Å as greater values add insignificant information but evoke significant difficulties during training [Kryshtafovych et al., 2011]. These data were later on used as foundation to design the global and local quality assessment methodologies which should in the end be able to predict the occurring deviation between corresponding atoms based solely on the presented model without knowledge of the experimentally determined structure.

The strategy was adopted for the CASP9 run [Moult et al., 2011], creating a dataset that can be used for evaluation of the trained model on a set of proteins whose fold is not definitely part of the training dataset.

## 5.2 Extension of BioJava

Implementation was solely done in Java, using as many existing libraries as possible. Especially BioJava [Holland et al., 2008; Prlic et al., 2012] proved useful due to excellent capabilities, when it comes to the initial parsing of PDB files and as a scaffold to develop necessary data structures. BioJava uses *Structure* objects to represent e.g. proteins which consist of one or more *Chains*, itself containing the actual residues represented as *Groups*. Each group features an individual property-map which enables the user to

store virtually any data directly linked to a certain residue. The property-map is a *Map* associated to each *Group*, making arbitrary *Objects* storable and allow the user to retrieve them by a defined key. Thereby, calculated energy values, residue descriptors as well as the observed deviation from the native structure could be stored. This data structure facilitated e.g. the computation of correlations between stored values and ensured that parsed or calculated values were linked to the correct residue. Last but not least, individual *Groups* could be fetched and handed to certain functions which then used the stored values to compute new values and add the result itself to the property-map.

As values in the *Map* vary in their value range, they are not directly comparable. However, they can be standardized by Z-scores. This results in a centering of all values around 0 and values deviating from the expected value are scaled by the number of SD they differ from that value. By a Java class, values can be normalized by Z-scores or with the knowledge of minimum and maximum values, preparing them for further processing. All entries of the property-map of a structure can be smoothed by a flexible window size  $N$ , which results in sequentially neighbored residues influencing each other in their values. A bell-shaped curve was implemented, resulting in the central residue having the biggest weight  $w_i$  while the influence of marginal ones vanishes.

$$w_i = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{i-\mu}{\sigma}\right)^2\right], \quad \sigma = \sqrt{\frac{N}{2}}, \quad \mu = \left\lfloor \frac{N}{2} \right\rfloor \quad (5.1)$$

While BioJava provides the property-map on residue-level, the *Structure* interface lacks this feature, rendering the storage of global values cumbersome. Thus, the wrapper class *Protein* was set up, accommodating exactly one structure while providing an analogue of the property-map as well as implementing the *Serializable* and *Comparable* interface which will prove useful later on.

Furthermore, several classes were created, which provide commonly used functions. E.g., one method performs the frequent, yet time-consuming cast of multiple variables stored in a *List* to an array of values. This was needed as the used statistics library StatUtils only processes double arrays. Others functions help copying files as well as compressing or extracting archives.

Last but not least, *StructureIOHandler* writes comma-separated values (**CSV**) files containing residue-wise values of specified properties. JFreeChart can be used to show line graphs directly or call overloaded functions to save the plots in **PNG** (portable network graphics) or **SVG** (scalable vector graphics) format to the file system. For three-dimensional visualization, a local PyMOL installation can be called and the user can execute any PyMOL command directly from a Java program. A function generates a PDB file by manipulating the column of the B-factor to contain one specified property. These files can then be parsed by PyMOL and it is possible to render high-resolution photos fit for publication automatically - e.g. visualizing a certain feature by exploiting the B-factor coloring.

All in all, a library evolved which made BioJava more convenient to use and add features necessary for upcoming tasks.

Based on this, the *EProfile* class was rewritten, now adding all values arising during computation as well as the final solvation and contact energy values to the property-map of each residue of a structure. Support for custom representation schemes was added, determining whether to represent residues by their  $C_\alpha$  atom,  $C_\beta$  atom, the last heavy atom of the side chain or the centroid of all atoms. Individual interaction distances can be specified and the necessary sequence separation can also be manipulated - default values are used for all energy calculations though.

For each mentioned MQAP a class was created to evaluate a PDB file by the according Web service or local installation and to receive the corresponding output file as well as to parse said file later on and add all information to the property-map of the appropriate residue.

Correlations were calculated between quantities computed by the MQAP, values of the energy profile approach and errors on residue-level which are assessed by the distance between equivalent atom pairs for single models of the CASP datasets. Line plots as well as colored PDB structures facilitated further studying coherences.

### 5.3 Training of the Evaluation Model

The GDT of each models is included in the CASP data as global evaluation measure and should be predicted with help of several global descriptors of the protein structure. On residue level, the distance between pairs of corresponding atoms should be forecast. As for now, both approaches are decoupled, but it would be possible to use the global assessment for more precise local quality values and vice versa.

Several features were defined which showed weak to mediocre correlations to the values to be approximated. For the local quality assessment the solvation and contact energies from the energy profile model were included. Furthermore, for each residue the number of long-range interactions was counted, which were defined as two residues being less than 10 Å apart but exhibiting a sequence separation of more than 8 positions. They were more closely assessed in form of amino acid-specific Z-scores, since each type of amino acid has a preference about the degree of fullness of its environment. Residues with a prejudicial Z-score below -0.5 were labeled as bad, whereas such with laudable scores exceeding +0.5 were labeled as good. Both numbers of other residues within a 10 Å sphere showing either label were computed. In accordance to QMEAN's approach [Benkert et al., 2011, 2008], the agreement between the by eGOR predicted and the actually computed, yet discretized energy profile was regarded by a boolean variable. Inspired by QMEAN, the relative ASA (divided the amino acid-specific maximum) and the loop fraction were also used as features. DSSP [Joosten et al., 2011; Kabsch and Sander, 1983] was used for calculation of the raw ASA values and to assign the secondary structure. The raw value of the loop fraction was 0 if a regular secondary structure element, in form of  $3_{10}$ -,  $\alpha$ -,  $\Pi$ -helix, beta sheet or bulge is present at this particular position, and else is 1. All features were smoothed by the previously described bell-shaped weighting rule and a windows size of 9, both the solvation and



contact energy term were kept also as raw values though, preventing the loss of too much information on residue level by smoothing. Thresholds for the Z-scores or the distance cutoff for long-range interactions were determined by varying said values to maximize their Spearman's rank correlation coefficient regarding the distance between equivalent atom pairs.

Roughly the same values were used as global descriptors of the structure. The average values for the solvation energy, number of long-range interactions, good and bad labeled residues, ASA, loop fraction and agreement between eGOR prediction and the calculated energy profile were utilized. However, all contact energy terms were summed up and divided by the square of the number of observed contacts within the structure. Last but not least, the dScore further characterized the agreement between both energy profiles on a global level. These features should now be used to predict the parsed GDT of each structure in the CASP10 dataset as global quality measure. Furthermore, all protein contained in the dataset of high-resolution structures and all experimentally determined structures of the CASP targets were included and their GDT was set to 100.0 as they are in fact the native structure.

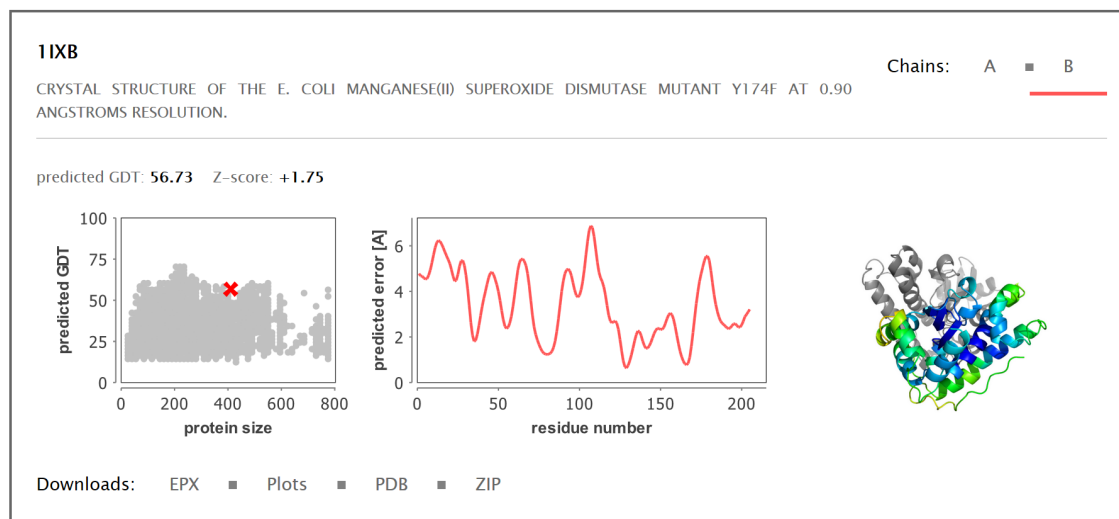
The Weka framework [Frank et al., 2004; Hall et al., 2009; Holmes et al., 1994] was utilized for inspection of the data and training of the assessment models. The random subspace method [Bryll et al., 2003; Ho, 1998] was chosen to realize the regression since it outperformed other techniques such as linear regression models, neural nets and conventional decision trees. The random subspace method randomly trims the initial feature space, creating a set of subspaces including a varying selection of features. For each subspace a classifier is individually trained and their predictions are combined e.g. by majority voting to obtain the final classification. Due to reduced dimensionality in the subspace even small sample sizes enable successful learning. On the other hand redundant features are eliminated [Panov and Dzeroski, 2007].

In both cases 10-fold cross-validation was employed, the subspace size was set to 0.9 and 200 iterations were done, resulting in a correlation coefficient of 0.87 for the global assessment and 0.52 for the local quality prediction. This resulted in two serialized models for local and global assessment respectively. Since the model files are several MB in size, their deserialization is realized by a static singleton ensuring that the models are only loaded once, even when facing concurrent or subsequent method calls. Two Java classes wrap the models and contain a method which hands all necessary data to the Weka models and assigns the local error to the property-map of the currently processed residue and global score to the structure respectively.

## 5.4 Deployment as Web Service

**CAMEO** (continuous automated model evaluation) is somewhat similar to CASP since it aims at assessing the quality of protein modeling approaches. However, participating teams do not have to submit their models to CAMEO, but provide a pipeline which is

then automatically queried when new targets emerge. Just like CASP, CAMEO does not strictly focus on the modeling category, but also pays attention to adjacent fields like ligand binding and model quality estimation. So I opted for developing a draft of a Web service to potentially participate and compare the quality of the here discussed approach with other state of the art methods.



**Figure 5.2:** Sample output for 1IXB

Each protein is described briefly by its PDB title. The global evaluation scores are accessible as numerical values and visualized by the scatter plot of all predictions. The local quality is presented as line plot and as colored structure rendered by PyMOL. By the chain selector in the upper right corner, the quality of individual chains can be inspected closely as figures will be replaced with the specified ones using jQuery to modify several CSS attributes. Certain subsets of the produced files can be downloaded.

Eclipse IDE for Java EE Developers was used to set up a Web service using JavaServer Pages (**JSP**) utilizing Apache Tomcat. The JSP provides the basic page structure featuring a form which allows the user to submit a file to the server. In the back-end a servlet handles the file upload and creates an instance of the surprisingly appropriate named *Job* class which stores information the user submitted while also being the vehicle to serialize all necessary data and making previous jobs retrievable by a link. Therefore, a universally unique identifier (**UUID**) allocated to each *Job*. Whether a single PDB file or an archive containing multiple structures was uploaded, is detected and any archive file is decompressed and unpacked. For each detected PDB file, the energy profile is calculated and several subroutines gather data which is needed to locally and globally evaluate the structure by the trained models - e.g. the secondary structure annotation by DSSP. A CSV file is created, containing the local error of each residue as well as all data which led to this score and some basic information in form of residue number, amino acid and annotated secondary structure. Chain-specifically, the local error plots are rendered by JFreeChart and the residue-wise error within the structure is visualized by PyMOL - again exploiting the B-factor coloring (Figure 5.2). The global assessment is presented by a scatter plot containing all predictions on the set used for training plotted

against the protein size; this enables the user to interpret the global assessment since this score strongly depends on the number of residues. To address this dependency, the score is also presented as Z-score, whereby it is compared to all other structures from the training dataset which have a size of  $\pm 10\%$ . Subsets as well as all created files are packed and compressed as archive files to create a convenient way of downloading all results. Finally, the *Job* instance is serialized to the file system and the user's browser is redirected to the result page.

For each *Protein* entry in the deserialized *Job* a container is created: stating filename, the structure's title from the PDB header, the global assessment scatter plot and both local deviation images for the first chain. A small JavaScript for chain selection implements the possibility to cycle through multiple chains (if present), replacing all plots with that of the chain the user is interested in. When multiple structures were processed, the *Comparable* interface allows sorting by the predicted global assessment Z-score, that way the highest quality model is displayed first and the reliability subsequently decreases when scrolling down.

It is also notable, that the routines realizing the local as well as global quality assessment are completely uncoupled from other back-end functionality, such as the energy profile calculation and generation of the output plots and figures or the front-end delivering the Web pages. Thus, all parts are interchangeable, a better trained model can be seamlessly deployed to the Web service. It is even imaginable to give the user the option of choosing among several different quality assessment routines.

## 5.5 Calculation of Consensus Energy Profiles

Several well-performing modeling pipelines or MQAP are clustering methods combining features of a preferably huge number of models, varying in quality. They use their diversity to formulate the consensus following the assumptions that often made predictions are likely to be correct. They can model protein structures or assess their quality by function as meta-predictors: e.g. some query a number of established modeling pipelines and retrieve their prediction and then use the combined knowledge of all other methods to make their own sophisticated prediction [Kryshtafovych et al., 2014a].

Is the same strategy applicable for energy profiles? Can a number of energy profiles originating from modeled structures be used to predict the profile of the native structure? Therefore, energy profiles will be computed, arranged and the mean energy at each position yields the consensus energy profile. The usefulness of the strategy was finally tested on the CASP10 dataset, whereby all models were used to calculate the consensus profile, and on the other side eGOR was used to predict the energy profile based on the sequence of the native structure. Both profiles were finally compared with the one computed, directly using the native structure and the distance was measured using the dScore.



## 6 Results

The mentioned MQAP were compared in their predications. Interesting factors are also the initial correlations of residue descriptors and how the model performed when compared to the existing methods. Furthermore, it should be able to perform equivalently well on independent datasets.

### 6.1 Correlations between CASP10 Data and Local Residue Descriptors

The downsized CASP10 dataset was analyzed by checking how well the several previously formulated residue features correlate with the distance between equivalent atom pairs. High correlations make any regression based on them more easy and reliable. At first glance weak to moderate correlations can be observed (Table 6.1). All but the

**Table 6.1:** Spearman's rank correlation coefficient for the EP-based approach on the CASP10 dataset

feature	correlation to $C_{\alpha}$ - $C_{\alpha}$ distance
solvation energy	0.35
solvation energy raw	0.22
contact energy	0.25
contact energy raw	0.13
interactions	-0.37
good interactions	0.31
bad interactions	-0.37
relative ASA	0.38
loop fraction	0.24
energy profile agreement	-0.13
EP-based composite score	0.60

values denoted as raw are smoothed by the bell-shaped weighting rule over a window size of 9, since this strategy improved the correlations significantly. Despite the loss of information on a particular residue, data from the surrounding residues is incorporated. Regarding the energy values, smoothing almost doubles the Spearman values, especially the solvation energy term correlates moderately with the  $C_{\alpha}$ - $C_{\alpha}$  distance of atom pairs of the superimposed structures of the CASP10 dataset. The number of long-range interactions and the number of residues in the environment falling below the threshold exhibit values of -0.37, but it is to be suspected that all three interaction values cover more or less the same informational content. As proposed by QMEAN's literature

**Table 6.2:** Spearman's rank correlation coefficient for QMEAN

feature	correlation to $C_\alpha$ - $C_\alpha$ distance
$C_\beta$ interaction energy	0.23
short-range $C_\beta$ interaction energy	0.14
all atom interaction energy	0.27
short-range all atom interaction energy	0.32
torsion energy	0.17
solvation energy	0.18
relative ASA	0.38
loop fraction	0.24
SSE agreement	-0.19
ACC agreement	-0.32
QMEAN composite score	0.49

[Benkert et al., 2011, 2008] the relative ASA and loop fraction are quite trivial features but correlate fairly well with the value to be approximated and therefore were adapted for the EP-based approach. The agreement term between the observed discretized energy profile and the prediction by eGOR does not seem to correlate exceptionally well, but in fact it adds essential information to the model which was not covered by other features. Omitting this descriptor results in lower overall quality of any trained model.

The combination of all values by the random subspace yields a regression coefficient of 0.60 which is a decent values and does not seem improvable by a big margin with the current knowledge [Kryshtafovych et al., 2014a].

QMEAN was studied in more detail (Table 6.2), as it is exceptionally well documented and its output provides not only the final score but also all terms leading to that exact result, making it quite easy to spot beneficial residue descriptors and understand its strategy better. Overall QMEAN's energy terms correlate slightly worse than that formulated by the energy profile approach. QMEAN's solvation energy term reaches a value of 0.18, whereas the energy profile one amounts 0.35, implying an overall better designed model for interactions with the solvent. In terms of the contact or interaction energies QMEAN outperforms the term calculated by the energy profile. As mentioned, QMEAN's calculation for the relative ASA and loop fraction was implemented in the EP-based approach, resulting in the values being nearly the same, thus showing the same correlation coefficients for both MQAP. QMEAN's agreement terms for the predicted secondary structure as well as ASA correlate with -0.19 and -0.32 respectively, and adding these features to the regression by the random subspace method improved the results slightly. However, integrating these values would also mean adding further dependencies on the prediction algorithms (namely ACCpro [Pollastri et al., 2002] and PSIPRED [McGuffin et al., 2000]) to the project and therefore I opted to ignore them for now. A regression combining all features of both approaches, exhibits a correlation co-

efficient of 0.65, implying that even though the EP-based approach was inspired heavily by QMEAN's methodology and they share much common ground, they cover diverse features and spot varying problems in protein structures.

The composite score yielded by linear combination of the single features shows a coefficient of 0.49, a lower than the one of the EP-based approach. However, the EP-based approach was designed and trained directly on this very dataset, meaning there is a huge bias towards the EP-based approach. In fact it also states that QMEAN was capable of generalizing its derived knowledge and making it applicable for other problems.

Both methods share a Spearman value of 0.71 for this particular dataset, attesting strong similarity.

## 6.2 Comparison of the MQAP

The local evaluation from all MQAP was computed for the dataset of high-resolution structures, arranged and correlations among them were quantified using Spearman's rank correlation coefficient (Table 6.3).

Mostly, vanishing Spearman values close to 0 can be observed. Weak correlations of

**Table 6.3:** MQAP local evaluation correlation matrix

	Verify3D	PROCHECK	VADAR	PROSESS	ProSA	ANOLEA	QMEAN	EP-based
Verify3D		-0.02	0.20	-0.03	0.02	-0.18	-0.28	-0.33
PROCHECK			-0.10	0.26	0.01	0.02	0.09	0.09
VADAR				-0.16	-0.17	0.11	-0.06	0.01
PROSESS					0.03	0.12	0.17	0.16
ProSA						-0.02	0.14	0.03
ANOLEA							0.62	0.62
QMEAN								0.69
EP-based								

0.20 can be spotted for the pairs of Verify3D and VADAR as well as 0.26 for PROCHECK and PROSESS. Verify3D's profiling approach shares a correlation coefficient of -0.28 with QMEAN and -0.33 with the EP-based approach respectively.

One would suspect rather strong correlations between similar approaches like the KBP or methodologies combining previously existing tools. However, a coefficient so close to 0 indicates contradictory statements, which is quite surprising as one would ponder to find at least some common ground when assessing the structural quality of residues. As Verify3D is directly part of VADAR and PROSESS is VADAR's successor, values

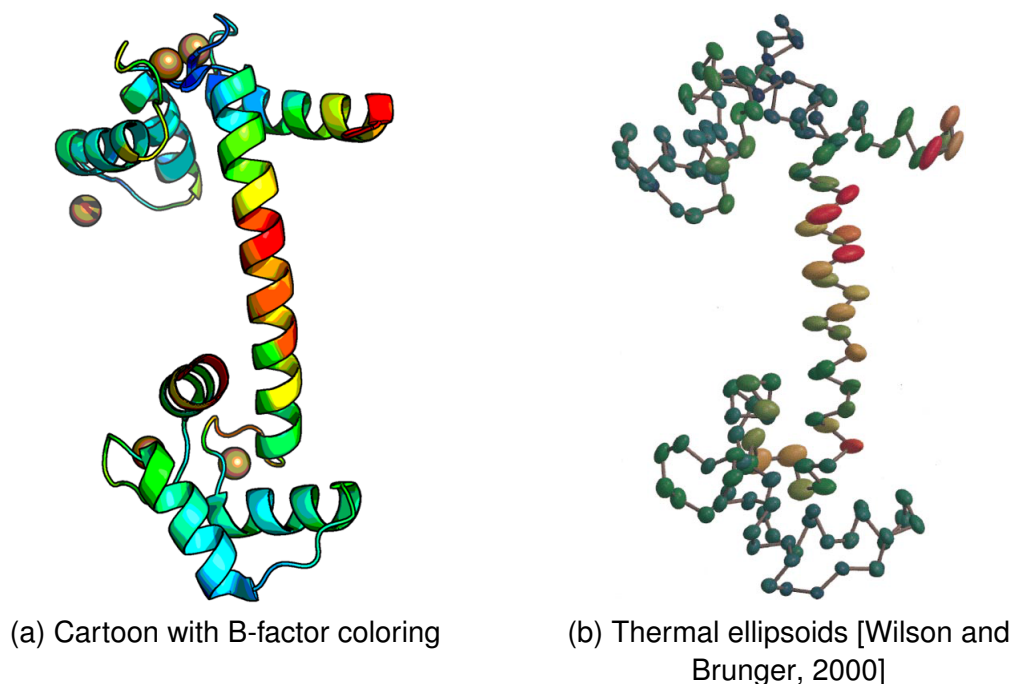
around 0.2 are achieved. To the same effect, Verify3D and PROSESS should exhibit at least some weak accordance. Why ANOLEA and QMEAN are that congruent in their assessment remains unclear; both are potentials of mean force, but nevertheless ANOLEA uses individual atoms to assign energy values whereas QMEAN considers whole residues. When these both tools are that similar, they should also correlate to ProSA - their mutual foundation. In general, ProSA shows outright low similarity to other methods. When summing up the single values for each MQAP, the QMEAN shows the biggest overall correlation towards all competitors, implying it is closest to the consensus predication of all methods.

Interestingly, the dataset of high-resolution structures also contains examples of protein disorder, enabling precise studying of the peculiar regions and giving a chance to compare MQAP on everyday tasks. Since the tools use various value ranges to quantify their assessment, Z-scores were used to standardize the data. Furthermore, regarding Verify3D and VADAR high scores indicate high quality, while for all other tools high scores indicate low reliability. Z-scores were calculated using observed values from the dataset of high-resolution structures. The consensus predication was derived by calculating the mean Z-score among all established MQAP (but not the EP-based approach). All results were visually inspected by rendering the structures in PyMOL while applying B-factor coloring according to the evaluation Z-score of a particular residue. A color range from red (2 or more SD worse than the expected value) to blue (2 or more SD better) was chosen.

Calmodulin (1EXR) was one of the structures which were inspected more closely: it binds to over 100 targets and is a major player in many cellular processes by acting as a secondary messenger. E.g. it is bound to Actin mediating contractions of smooth muscle tissue. The high flexibility and plasticity of 36 residues located in the linker helix connecting four  $\text{Ca}^{2+}$  binding sites are crucial in mediating the protein's function (Figure 6.1). Many different conformations can be observed depending on the present binding partner [Kovalevskaya et al., 2013; Wilson and Brunger, 2000]. In fact, the exact orientation of both domains is almost random in solution and furthermore, methionine residues governing interactions with other molecules also show decent disorder [Chou et al., 2001].

Because of the high disorder in the linker region, it is prone to high predicted uncertainty among MQAP (Figure 6.2). Visual inspection shows that Verify3D, PROCHECK, QMEAN and the EP-based approach evaluate the suspicious region as unreliable. Verify3D tends to get negligent when the linker sequence merges into the more ordered parts of the protein. Furthermore, Verify3D as well as the EP-based approach evaluate regions in the  $\text{Ca}^{2+}$  binding domains in the same fashion as the linker helix, especially when said regions are exposed to the solvent. VADAR, PROSESS, ProSA and ANOLEA do not characterize the linker region and the  $\text{Ca}^{2+}$  binding EF-hands differently, in fact they all show little variance in its assessment along the sequence overall. They differ in the average quality they assign to the protein, but overall they do not draw attention





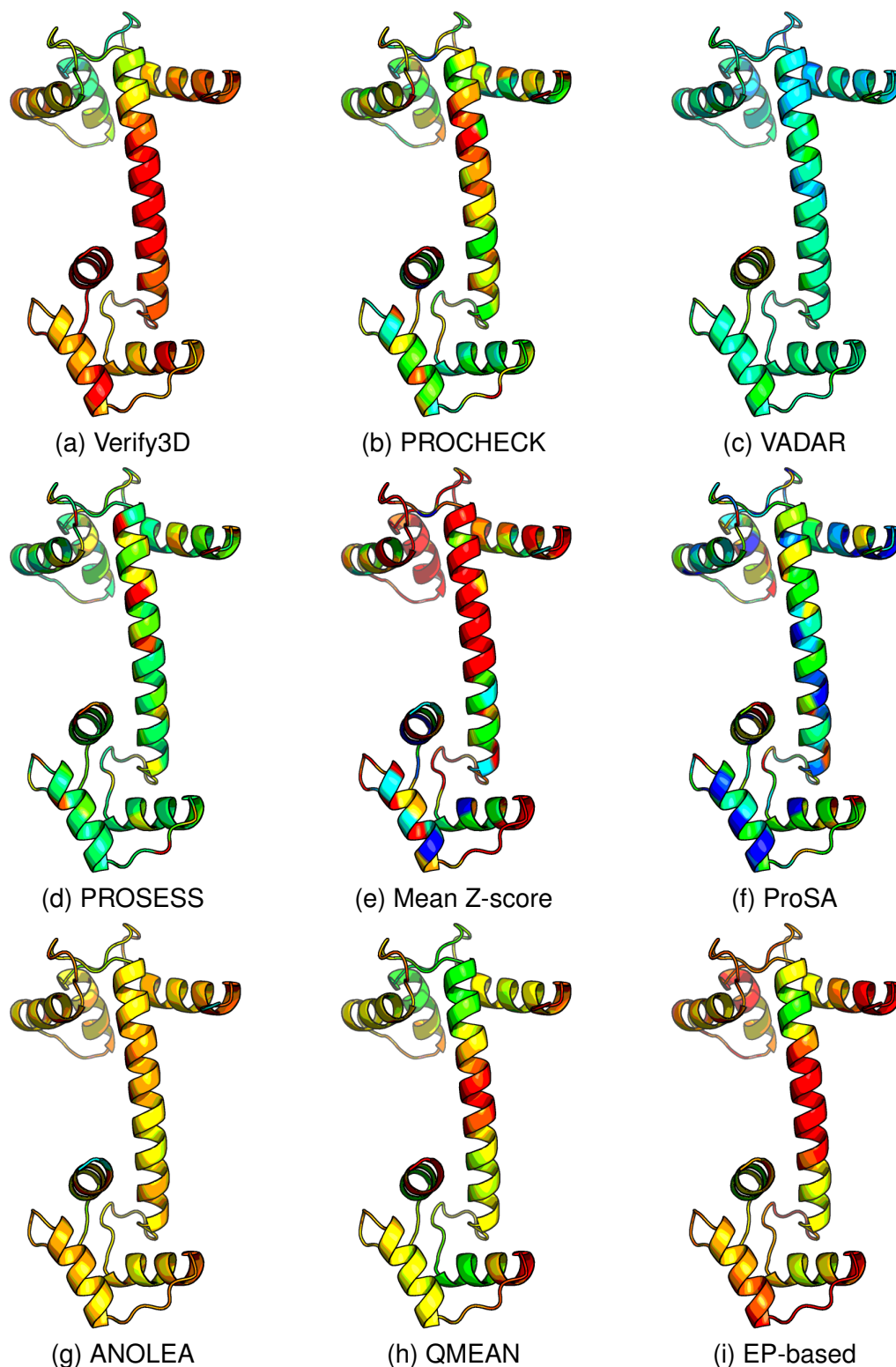
**Figure 6.1:** Structure visualization for Calmodulin (1EXR)

(a) 1EXR binds  $\text{Ca}^{2+}$  ions (ochre spheres) and features a linker helix of high disorder which connect both ion binding domains. Temperature factors were visualized using a range from 5 (blue) to  $25 \text{ \AA}^2$  (red). (b) The ellipsoids characterize the position of  $C_\alpha$  atoms at 90% probability level. Large deviations occur in the central helix, the hydrophobic binding sites and the termini [Wilson and Brunger, 2000]. MQAP should be able to spot the peculiar regions exhibiting high B-factors.

to the linker helix. Neither of them captures the ambivalent properties of the structure. Even though four tools exhibit a quite constant assessment and the other tools evaluate the linker helix worse than the rest of the protein, combining all scores by a Z-scores leads to an assignment of overall bad quality. The partially successful differentiation between both protein parts is lost. Due to the nature of energy-based approaches to be depended on the protein size, smaller structure get punished unjustifiably and the assessment of all but ProSA results in being rather unflattering.

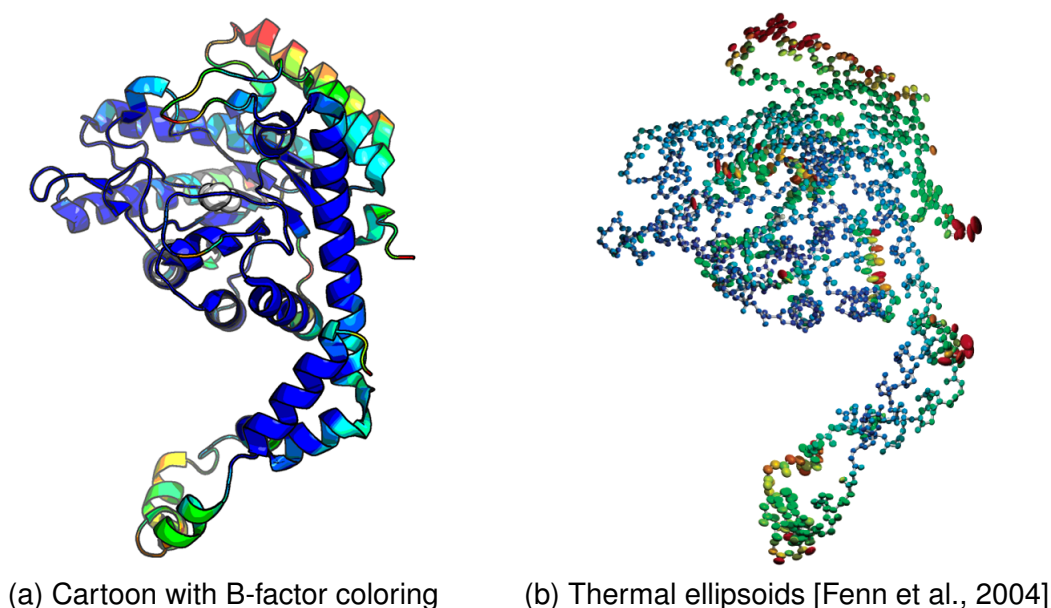
Xylose Isomerase (1MUW) is also notable, an enzyme promoting the conversion of several aldose and ketose sugars. Obviously, it is an essential part of the sugar metabolism, and furthermore of industrial interest for the production of high fructose corn syrup. The protein was studied in conjunction with glucose (1S5M) and xylitol (1S5N) and one  $\text{Mn}^{2+}$  ion adopts different positions depending on the substrate located in the active site. In consequence side chains involved in metal ion coordination are prone to some disorder as well, but in total the TIM barrel region appears well-ordered. Furthermore, the top loop in the foreground (residues 22-27) and a major part of the upper right  $\alpha$ -helix (residues 60-80) feature significant directional displacement (Figure 6.3) [Fenn et al., 2004].

Verify3D labels especially the isolated part as unreliable. However, the distorted coil



**Figure 6.2:** Local evaluation for Calmodulin (1EXR)

1EXR's central linker helix shows exceptionally high flexibility over 36 residues [Wilson and Brunger, 2000]. The output of each MQAP was standardized by Z-scores, since the tools use varying ranges and regarding Verify3D and VADAR, high scores indicate high quality. Furthermore, the mean predication of all MQAP was computed by averaging all their standardized values. For coloring, a range from -2.0 (blue, better quality than average) to +2.0 SD (red, worse) was applied. Verify3D, PROCHECK, QMEAN and the EP-based model evaluate the peculiar part as low quality, although only PROCHECK can reliably distinguish the linker region and the  $\text{Ca}^{2+}$  binding parts.

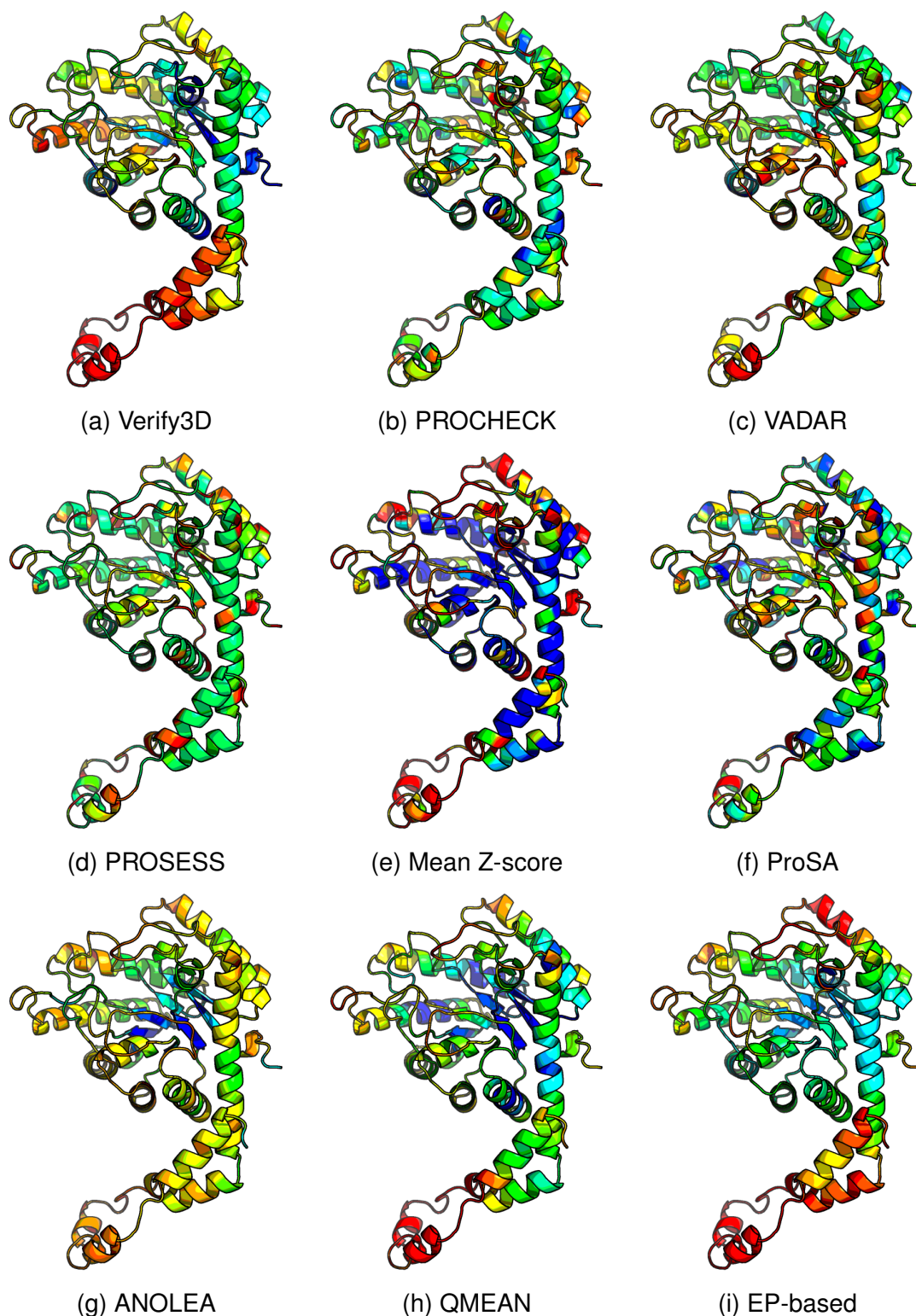


**Figure 6.3:** Structure visualization for Xylose Isomerase (1MUW)

(a) 1MUW converts aldose and ketose sugars. The silver spheres represent catalytically active  $\text{Mn}^{2+}$  ions embedded in a TIM barrel. Temperature factors were visualized using a range from 5 (blue) to 20  $\text{\AA}^2$  (red) [Fenn et al., 2004]. (b) The ellipsoids characterize the position of  $C_{\alpha}$  atoms at 50% probability level. Large deviations occur in the exposed, isolated region in the lower left, the termini and the helix in the top right of the figure [Fenn et al., 2004]. MQAP should be able to spot the peculiar regions exhibiting high B-factors.

and helix are evaluated similar to the rest of the TIM barrel. Even though PROCHECK dealt exceptionally well with the previous task, this time it cannot point out the peculiar parts of the protein. Similar behavior also shows PROSESS and, yet again, the variance of the assigned scores is low. However, the disordered coil and helix as well as the isolated chain fragment exhibit slightly worse scores. Just like Verify3D, VADAR is capable of spotting problems in the region in the bottom left of the structure but misses the other suspicious fragments. Again ProSA fails to spot the suspicious parts or to distinguish residues exposed to the solvent and such buried in the hydrophobic core of the protein, the whole assessment is checkered and seems quite random. Last but not least, ANOLEA, QMEAN and the EP-based approach manage to spot the protein parts with high fluctuation and simultaneously attest the substantial quality of the TIM barrel. Their exact assessment differs slightly though. While the previous Z-score analysis for 1EXR showed overall unfavorable predications of the MQAP, this time the structure is evaluated overall as relatively good. The tools are in accordance and assign bad quality measures to the disordered helix as well as the part located far away from the main part of the protein. Furthermore, the termini plus several coil regions facing the solvent receive unflattering scores.

The global predications of the MQAP could not be reasonably compared since VADAR does not provide any composite score and according to literature Verify3D provides a



**Figure 6.4:** Local evaluation for Xylose Isomerase (1MUW)

1MUW features a disordered helix in the upper right and an isolated region located away from the main part of the protein. The output of each MQAP was standardized by Z-scores, since the MQAP use varying ranges and regarding Verify3D and VADAR, high scores indicate high quality. Furthermore, the mean predication of all MQAP was computed by averaging all their standardized values. For coloring, a range from -2.0 (blue, better quality than average) to +2.0 SD (red, worse) was applied. Verify3D, QMEAN and the EP-based model successfully label the peculiar parts as unreliable.

global score but in fact it is not part of the Web service's output.

### 6.3 Performance on CASP Datasets

Spearman's rank correlation coefficient shared by the actual GDT and the one predicted based on the model amounts 0.91 for the CASP10 dataset. Normalizing by the Z-score regarding proteins of similar size results in a slight decrease of the correlation coefficient to 0.90.

Visual inspection of some randomly chosen CASP10 models for target T0645 (Figure 6.5) indicates that the trained model is in agreement with the actual present error of each residue roughly all the time. There is a periodical alternation between regions with minimal deviations and parts where the model contains significant errors of more than 5 Å.

**Table 6.4:** Spearman's rank correlation coefficient for the EP-based approach on the CASP9 dataset

feature	correlation to $C_{\alpha}$ - $C_{\alpha}$ distance
solvation energy	0.38
solvation energy raw	0.22
contact energy	0.23
contact energy raw	0.09
interactions	-0.36
good interactions	0.25
bad interactions	-0.37
relative ASA	0.41
loop fraction	0.32
energy profile agreement	-0.21
EP-based composite score	0.47

The CASP9 dataset was used to assess the method's quality with protein structures which were not part of the dataset used for training. Even though not all models were used for training, all structures share a common fold especially as only somewhat similar models were selected due to the exclusion of models with a GDT below 20.

All in all, the EP-based approach performs worse on this (again downsized) dataset (Figure 6.4). Overall, the Spearman value amounts for 0.47 and almost all contributing residue descriptors exhibit correlation coefficients close to the ones calculated for the CASP10 dataset, indicating that the model is not capable of generalizing absolutely unbiased even though the informational content of the chosen features is almost the same. Only the loop fraction and the agreement term between the calculated energy profile and the prediction based on the protein sequence show stronger correlations. The mean loop fraction is with a value of 0.47 marginal higher than the one for the

CASP10 dataset, accounting for 0.44 - indicating that the structures part of the CASP9 run contain more coil in general, which will probably result in less correct structure predictions as well as lower attested quality. That is in fact the case as the average local evaluation is 4.08 and 3.75 for the CASP9 and CASP10 dataset. Usually, targets hard to model are also hard to evaluate precisely. Remarkable is also that QMEAN is not able to return results for all submitted models and in consequence models with missing data were excluded from the analysis. Interestingly, the EP-based approach performs even worse when considering all models and the correlation coefficient to the  $C_\alpha$ - $C_\alpha$  distance drops to a value of 0.40. Possibly, QMEAN performs way worse on the models for which no results were returned too.

For sakes of comparability, the correlations of the residue descriptors used by QMEAN were inspected in detail again (Table 6.5). Especially the terms for the  $C_\beta$  interaction energy as well as the one for all atoms show an increase of their correlation coefficients of more than 0.1. In contrast to the better correlation of the EP-based energy profile agreement term, QMEAN agreement features cannot capture the quality of a residue as well as before. The observations for the relative ASA and loop fraction are again similar to the values of the EP-based approach. Overall, QMEAN performs slightly better compared to the CASP10 dataset, in direct comparison QMEAN manages to outperform the EP-based approach, even though the Spearman values of the composite scores are approximately the same magnitude. It also remarkable that QMEAN is in fact only a trivial linear combination of the values of its residue descriptors, while the EP-based approach uses a quite sophisticated machine-learning approach.

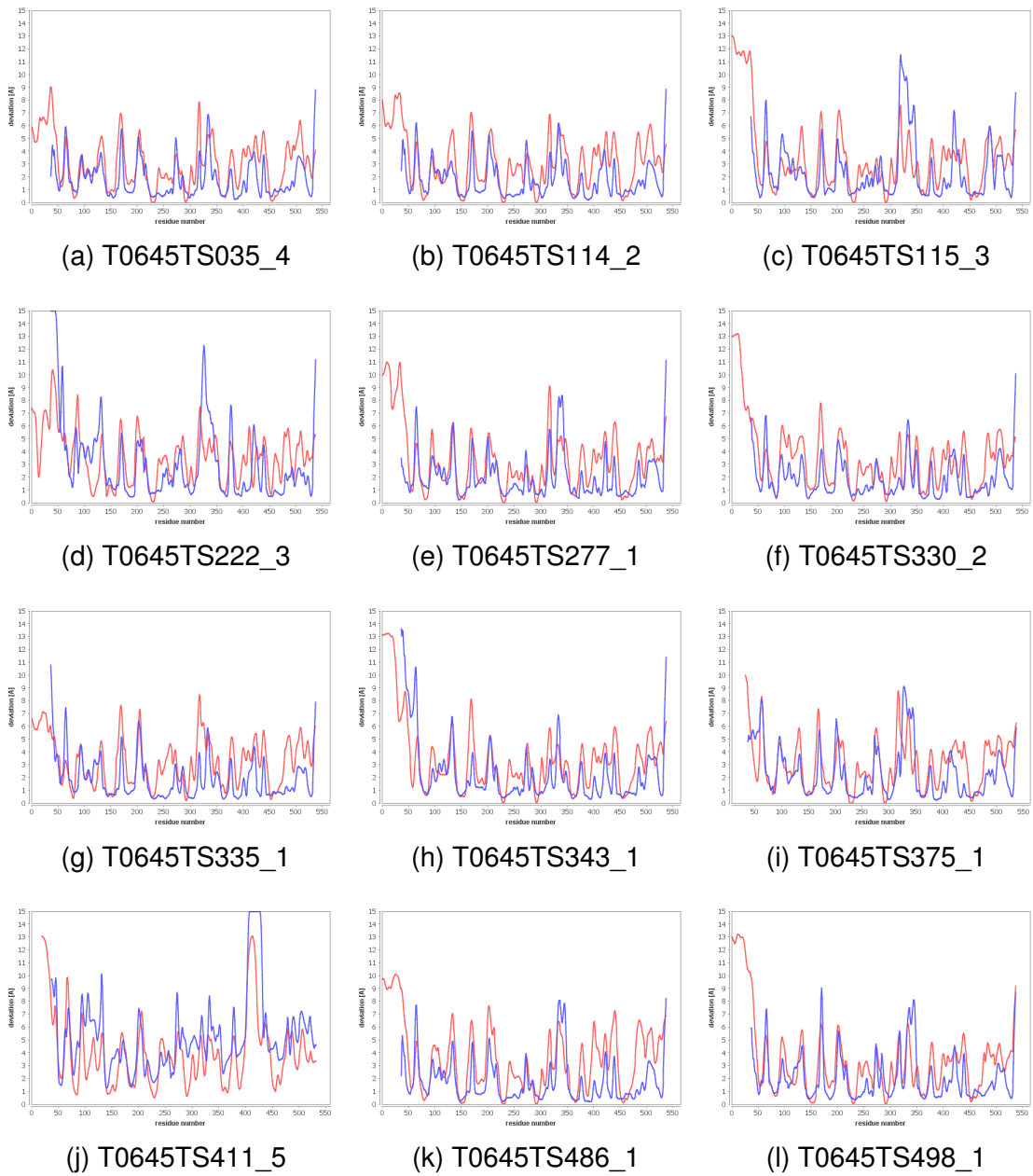
The predicted GDT and Z-score of the global quality assessment strategy correlates

**Table 6.5:** Spearman's rank correlation coefficient for QMEAN on the CASP9 dataset

feature	correlation to $C_\alpha$ - $C_\alpha$ distance
$C_\beta$ interaction energy	0.33
short-range $C_\beta$ interaction energy	0.17
all atom interaction energy	0.40
short-range all atom interaction energy	0.17
torsion energy	0.13
solvation energy	0.21
relative ASA	0.43
loop fraction	0.34
SSE agreement	-0.17
ACC agreement	-0.24
QMEAN composite score	0.53

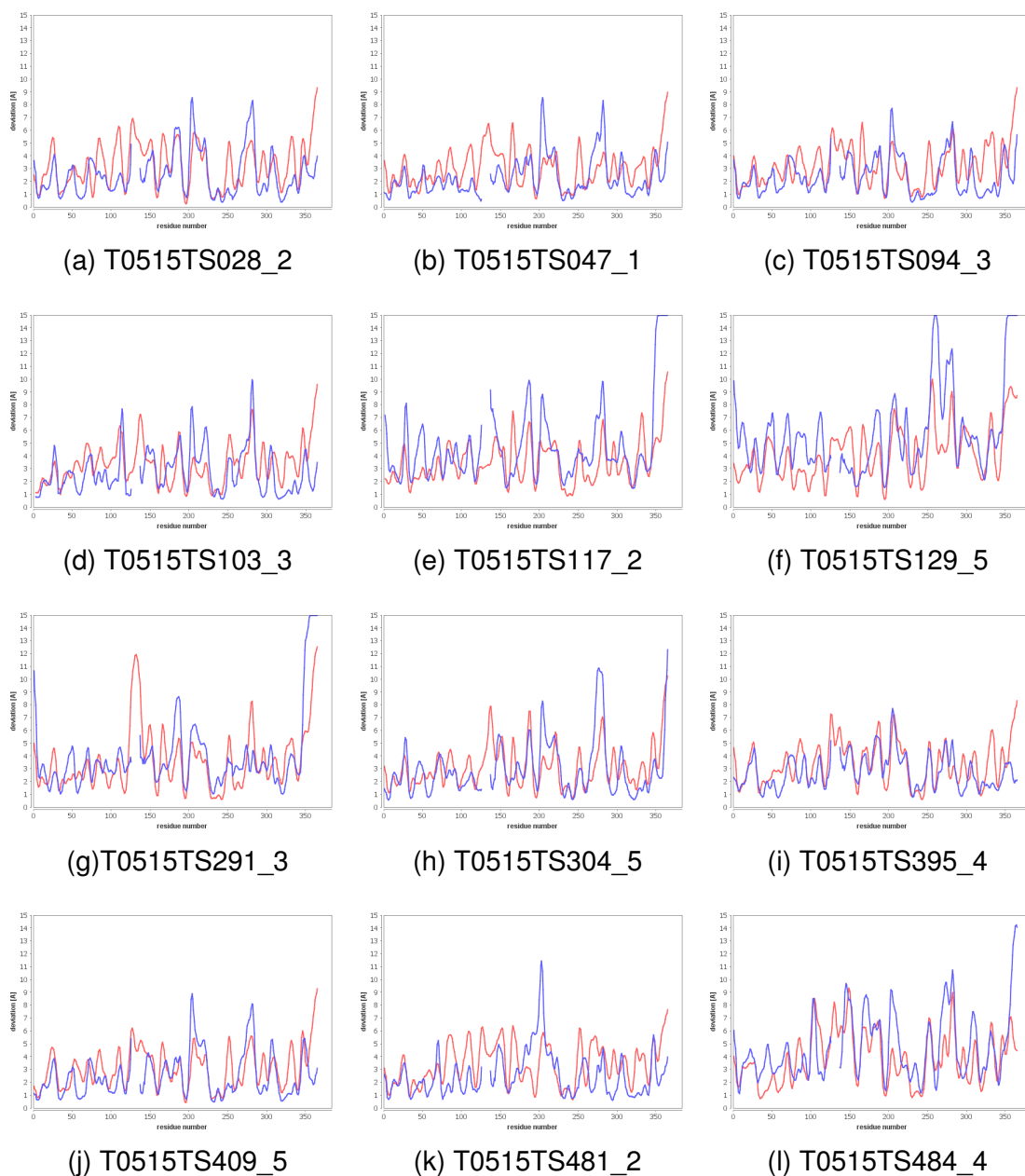
with 0.46, respectively 0.44, to the actual GDT of the structure alignments. The correlation coefficients are almost halved, implying lessened abilities of the model to predict the global quality of protein targets which were not used for training.





**Figure 6.5:** Variety of evaluations for CASP10 target T0645

12 models were selected randomly for target T0645 (PDB: 4F7A). The observed  $C_{\alpha}$ - $C_{\alpha}$  distance (blue) as well as the by the EP-based model predicted deviations (red) are plotted for each residue of the protein. The EP-based evaluation approach is able to predict the real deviations roughly correct and also uses the appropriate value range - meaning extraordinarily high distances get assigned extraordinarily high predictions. However, closer inspection shows minor differences for particular models such as *g* (T0645TS335\_1) or *k* (T0645TS486\_1). In general, the EP-based model tends to overestimate the real error slightly. Potentially up to 5 of the shown models could also be part of the downsized dataset used for training, but the behavior could be generalized and the EP-based approach can also handle models not part of the training dataset appropriately.



**Figure 6.6:** Variety of evaluations for CASP9 target T0515

12 models were selected randomly for target T0515 (PDB: 3MT1), whose fold not covered by the CASP10 dataset. The observed  $C_{\alpha}$ - $C_{\alpha}$  distance (blue) as well as the by the EP-based model predicted deviations (red) are plotted for each residue of the protein. Again, the plots exhibit the typical pattern of regions of small error followed by protein parts of lower quality and vice versa. However, the model has a hard time estimating the occurring distances correctly, e.g. the actual deviations are assessed false for models *b* T0515TS047\_1 and *c* T0515TS094\_3. In particular, the distances for the residues 200-230 are correctly approximated most of the time. The C-terminus is prone to serious deviations in both the aligned structures as well as regarding the prediction of the model. In summary, the approach can estimate the quality of individual residues of models, even when their fold was not presented during training.



Besides the numerical analysis of the predications of the EP-based model, they were also visually inspected for one target of the dataset. It was mandatory that the target could be considered an easy target which was overall well modeled, as it is pretty pointless to assess the local quality of a structure when the whole model is false and there are no minor errors to spot, but one large. Figure 6.5 shows the local error plots for target T0645 (PDB: 47FA). Usually, the local error plot of a protein chain periodically changes between parts of high quality and such exhibiting huge deviations regarding their  $C_{\alpha}$ - $C_{\alpha}$  distance. Unreliable regions are mostly located in coil regions between ordered secondary structure elements, which are exposed to the solvent. Thus, these parts of the protein enjoy a higher degree of freedom and are more flexible regarding their exact position. In consequence, structure alignments of proteins tend to differ heavily in such regions, even when the hydrophobic core of the protein can be aligned perfectly. Still though, the spatial localization of coil exposed to the solvent can be determined less exact by structure determination methods and protein structure prediction approaches struggle with the variety of possible conformations. The EP-based approach detects this pattern and assigns the same fluctuations. Also, the range of values of the models is predicted correctly, meaning models with overall low atom distances get assigned low predictions and vice versa. Due to the same assumptions concerning the flexibility of coil regions, the termini of a protein also tend to feature high errors. This propensity is spotted by the model. Even though the general aspects of the local errors can be predicted correctly, some erroneous regions remain unrecognized in detail and on the other side low quality is predicted where the modeled structure is in fact correct.

Because the target T0645 out of the CASP10 dataset was used to train the quality assessment model in the first place, one target of the CASP9 dataset was selected which is not of the same fold as any of the proteins used for training. The general view remains equivalent (Figure 6.6), but the disagreement between both plotted values seems to increase this time. More errors remain undetected and the model predicts more false positive errors. The evaluation method is able to handle novel folds in a suitable manner though.

## 6.4 Approximating Energy Profiles Using the Consensus Method

Last but not least, the capability of the consensus approach was analyzed. It should approximate the energy profile of a native structure without directly knowing the structure but only a number of predicted models. Some of these models are similar to the actual experimentally determined structure while others contain local errors, some are misfolded and there are also completely unfolded ones.

Despite these seemingly harsh conditions, the sequence-based eGOR methodology is outperformed by the consensus approach in all cases (Table 6.6). While the dScores for the consensus method ranges from 0.12 to 0.96, eGOR's prediction cannot approx-

imate the energy profile of the native structure as well, resulting in dScores between 1.55 and 3.20 with an average of 1.95. Thus, the consensus approach manages to extract the correct energy profile even when most used models are entirely wrong. The standard deviations of both variables are of the same magnitude, indicating similar dispersion of both measurands. The Spearman's rank correlation coefficient shared by both approaches amounts to -0.24, implying that there is no common ground between them and one method does not necessarily perform well when the other one does, and neither vice versa.

**Table 6.6:** Descriptive statistics for methods predicting the native structure's energy profile

method	mean	minimum	maximum	SD
consensus method	0.34	0.12	0.96	0.15
eGOR prediction	1.95	1.55	3.20	0.19

## 7 Discussion

Even though the exact approaches of QMEAN and the energy profiles differ, their energy models seem to capture related properties. It can be stated that the relative ASA and loop fraction may be trivial features but they indeed help the performance of a designed MQAP and should be covered in the most cases. Furthermore, it would be really appealing to study the individual features contributing to the composite score of each tool (just like it has been done for QMEAN) and study the exact impact of each variable, extracting the most suitable categories for model building. But unfortunately, tools like VADAR or PROSESS do not make information of that kind accessible. However, one could suspect that there are even more informative residue descriptors, which have not been spotted yet, whose incorporation could result in a slight improvement of the method's overall quality. Not the combination of the composite scores of the MQAP by averaging the Z-scores leads to an informative consensus but instead, the best features should be combined in an independent regression yielding a tool combining aspects of all methods - resulting in an indirect consensus.

All in all, a solid MQAP evolved, even though slight modifications could truly make it top-notch. Still, a more carefully designed training dataset containing a bigger number of unique protein folds could improve the quality which is achievable by the training process. Currently 10 CASP datasets are available. They could all be downloaded and their sequences and experimentally determined structure could be aligned against each other, enabling selection of a number of targets covering all present folds in the CASP data without being biased towards any particular fold. A small number of suitable representatives could be selected, analogous to the previous strategy randomly or preferably again by clustering. Training based on this dataset would cover roughly 500,000 models shrank into manageable size. Since especially regression methods acting like a black box, where one cannot understand the actually happening trainings process in detail, are prone to overfitting: only covering around 100 targets from one CASP run could have devastating consequences. The local quality assessment routine seems to be able to generalize enough to be applicable for other problems such as particular structures of the data of high-resolution structures or the older CASP9 dataset. In contrast, the significant decrease of capability of the global assessment model to approximate the overall quality implies learning by heart, of which combination of attributes leads to which GDT. Without actually spotting which arrangements of features leads to which measured global quality. At least the global assessment arc needs to be redesigned using this more appropriate dataset, but the local assessment method should follow contemporaneously for sakes of uniformity. Also the results of the local assessment - especially in form of the mean assigned local score - seem to correlate well with the overall quality of a model, thus incorporating them should lower the error of a newly trained model for the global protein structure evaluation.

The observations for the consensus energy profile indicate that they are capable of ap-

proximating the energy profile of the native structure way better than eGOR and could be a useful foundation for developing an energy profile-based quality assessment program for ensembles of models. However, as for all consensus methods participating in the CASP experiment it is questionable what the minimal number of processed models, to still reliably predict the energy profile of the native structure or at least perform significantly better than the merely sequence-based eGOR algorithm.

Regarding the dataset of high-resolution structures, Verify3D performs well when visually inspected. According to the Spearman's rank correlation coefficient, it is also similar to VADAR which incorporates Verify3D's profiling approach and all energetic MQAP except ProSA. In fact, the profiling idea also formulates an environment for each amino acid that it is probably located in. PROSESS uses maybe too many other features resulting in Verify3D having a too low impact resulting in almost no correlation between both programs.

PROCHECK captures the high disorder of the linker helix in 1EXR, but evaluates the quality of 1MUW rather uniformly, not spotting peculiar regions. Surprisingly both, VADAR and PROSESS, showed underwhelming results: they make a rather static statements on the overall quality of a structure and for individual residues this predisposition does seldom change. Both approaches try to combine many different feature categories, but indeed they seem to perform worse than the contributing methods like Verify3D or rather simple formulated MQAP like ANOLEA.

One can suspect that even when PROCHECK, VADAR and PROSESS lack a holistic point view and cannot spot the major problems present in the structures, they still derive a much more fine-grained statement on stereochemical properties of single residues which cannot be captured by the energetic methods instead. Standard coarse-grained energy models will e.g. not detect clashing atoms. It seems reasonable to cover both main categories for assessing the quality of structures: plausible bond lengths as well as angles and the energetic aspects.

ProSA features assessment far away from the other tools and for particular problems ProSA fails to spot apparent problems or even to provide reasonable evaluations. The parser's arrangement of the read values to the correct residue may be faulty since ProSA's output file only contains the calculated energy values but no hints to which residue they exactly belong. However, no offset was observed, meaning all values in the file were assigned to the same number of residues of the *Structure* object. Possible errors could also occur due to the fact that the Windows version of the program was used, an operation system experiencing some serious antipathy by the bioinformatic community, perhaps this particular version is neglected and present bugs were not fixed or even spotted.

ANOLEA shows little variance in the assigned energy values for both 1EXR and 1MUW. However, regarding the Xylose Isomerase differentiation between the core of the TIM barrel and parts exposed to the solvent can be observed. It seems that the value range for this MQAP strongly depends on the processed protein and cannot be generalized for all structures. As energy model only regarding contacts between, residues it will

probably assign overall better scores to large structures with a small radius of gyration. Potentially calculated Z-scores should maybe not describe the general population of observed values, but only state whether a particular residue exhibits a better or worse score than the average value of that one structure, resulting in a more sensitive representation.

QMEAN and the EP-based approach are very similar since both use two identical values in form of the ASA and loop fraction. Both features had huge impact on the regression of either approach since they correlate exceptionally strong with the approximated variable. In consequence, both MQAP derive congruent overall statements, differing only slightly in the exactly assigned scores. As mainly energetic methods, QMEAN as well as the EP-based approach tend to put regions at a disadvantage, which are exposed to the solvent, even though no true errors are present.

It should also be mentioned that some MQAP smooth their predications, for some the user can specify the window size for averaging and some output the raw values. It seems that smoothing is advantageous as it lowers the influence of individual residues and takes the direct environment of a residue into account. One can suspect that equal smoothing of the predications of all MQAP would lead to stronger correlations among them or at least level the field. However, I used the standard output option of each MQAP, assessing the genuine predication of all tools - the values the authors wanted to provide the user by default.



## 8 Outlook

As for the methodology, correlations of some form of torsion energies or features close to the stereochemical approaches could be integrated. However, that would result in a decrease of the importance of the initial KBP. Furthermore, there seems to be some bias towards certain folds and problems when dealing with ones which were not covered by the training dataset. All available CASP datasets could be utilized, their targets clustered according to their sequence identity and structural similarity and then again a small number of representatives could be used to form a training dataset. Also, a modified version of the approach could be designed which is not based on one single model, but on a set of models. Such clustering approaches perform way better due to the increase of evaluable information. It is not yet understood what the minimal number of models is in order to still derive stable results - the latest CASP run indicates that around 20 models of varying quality are sufficient, making these approaches much more practicable than previously believed. Upcoming CASP experiments will probably feature a dedicated category providing a heavily reduced dataset [Kryshtafovych et al., 2014a]. This strategy should be adapted for membrane proteins as well, as they feature drastically different environmental conditions since they are embedded in a hydrophobic lipid bilayer. It is questionable whether occurring preferences for bond lengths, bond angles, ASA and especially energetic terms can be generalized for all proteins or even for the varying positions regarding the membrane protein topology. Even though the community is aware of the problem, no dedicated MQAP handles membrane proteins individually. The here presented protein energy profiling already pays attention to the altered conditions within membrane proteins, yet correct prediction of the protein topology would be mandatory. However, it could be challenging to obtain a sufficient number of membrane protein CASP targets for training since they are somewhat neglected by the CASP experiment.

Also, it would be reasonable to get rid of the dependencies on other programs where possible. E.g. DSSP occasionally produces hard to handle errors and enforces the local installation of the tool. BioJava has some rudimentary classes trying to provide functions for the calculation of the ASA and to assign secondary structures according to the rules used by DSSP, then again they are real buggy at the moment and it probably takes some time until these features are reliably usable. Potentially upcoming BioJava versions will also add methods to predict ASA and secondary structure elements the way ACCpro and PSIPRED do.

Furthermore, some minor tweaks of the tool are necessary, e.g. the Web interface allows the user to provide an email address and it is not yet implemented that the user will actually receive the results. However, this is essential for the registration in the CAMEO experiment. The Web interface could be improved by the use of HTML5 elements like <canvas>, JavaScript and WebGL to dynamically render all plots in the browser and even visualize structures colored by the predicted error, giving the user the option to ro-

tate and zoom into the protein structure. Unfortunately, libraries providing such features only start to emerge and struggle with the extremely limited resources and currently cannot provide advanced options like B-factor coloring or an appealing cartoon representation like provided by PyMOL.



## 9 Summary

Just like there is no innocuous way to determine the quality of a structure by the experimental data alone [Laskowski, 2005], there is also no MQAP spotting all errors that can occur in protein structure and quantify the local uncertainty correctly [Cozzetto et al., 2007]. A similar solution is applicable for both problems however: one can trust a number of independent, yet consistent protein structure models and one should use as many MQAP as possible to evaluate a questionable structure. Even though it does not seem practical to combine them directly by a Z-score, plain visual inspection can provide some valuable insight. Hereby, ideally tools correlating not at all should be gathered as they truly provide new information and not only make the same statement and spot the same problems as some other tools. Especially, the combination of stereochemically motivated MQAP like PROCHECK and KBP such as QMEAN or the EP-based approach seems reasonable as both complement each other [Kryshtafovych and Fidelis, 2009]. The best residue descriptors should be extracted and utilized for an independent regression resulting in the formulation of the indirect consensus of multiple MQAP. Generally speaking, it was shown that KBP correlate well with the occurring error in protein structure models and methodologies based on them perform better than mainly stereochemical approaches like PROCHECK, VADAR or PROSESS. Although the relative ASA is a quite simple and obvious residue descriptor, it correlates well with the measured local error of residues and is at least as important as any energy term.

This principle was applied for the design of the EP-based approach by adopting the relative ASA and loop fraction terms originating from QMEAN. This led to a MQAP which - at first glance - performs well on both the CASP datasets and the dataset of high-resolution structures with their peculiar proteins. As a method utilizing only one model to derive its statement, it is quite a rarity - only five of these programs participated in CASP10 [Kryshtafovych et al., 2014a]. For unbiased evaluation by a third party, the Web service draft should be topped off, made (privately) available on the Internet and be signed up for the CAMEO project. Thereby, protein structure models would be submitted to the EP-based approach and the quality of its predication monitored automatically. Furthermore, the MQAP would be compared with other state of the art methodologies and analyzed with the knowledge derived from almost a dozen of CASP runs.

Major advantage of the EP-based approach besides the actual formulated predication are little computation times. The actual calculation of the energy profile and assessment by both random subspace models usually takes less than a second. However, the rendering by PyMOL in order to present the results can be quite time-consuming, especially for huge structures. The proposed user interface was again adapted from QMEAN, however providing the plots as vector graphics and using ray-traced PyMOL visualization of the structure instead of the quite unappealing results obtainable using Jmol. In general, the interface of the EP-based approach should at least be on par with

the one of QMEAN which itself is unmatched among the presented tools. Even though emphasis should be on the actual predication of any MQAP, it is also important not to lose sight of usability, as convenient to use tools encourage users not only to employ that very tool but to possibly evaluate their structure in the first place. A locally runnable version of the tool could also be designed, providing users the same features in a graphical user interface or by command-line. Java is predestined for such tasks because of its platform independence and of its convenient way to create dynamically generated Web pages.

## Bibliography

- S. Altschul. Gapped BLAST and psi-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402, Sep 1997. ISSN 1362-4962. doi: 10.1093/nar/25.17.3389. URL <http://dx.doi.org/10.1093/nar/25.17.3389>.
- Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, Oct 1990. ISSN 0022-2836. doi: 10.1016/S0022-2836(05)80360-2. URL [http://dx.doi.org/10.1016/S0022-2836\(05\)80360-2](http://dx.doi.org/10.1016/S0022-2836(05)80360-2).
- C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, Jul 1973. ISSN 1095-9203. doi: 10.1126/science.181.4096.223. URL <http://dx.doi.org/10.1126/science.181.4096.223>.
- K. Arnold, L. Bordoli, J. Kopp, and T. Schwede. The SWISS-model workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, 22(2):195–201, Jan 2006. ISSN 1460-2059. doi: 10.1093/bioinformatics/bti770. URL <http://dx.doi.org/10.1093/bioinformatics/bti770>.
- P. Benkert, M. Kunzli, and T. Schwede. Qmean server for protein model quality estimation. *Nucleic Acids Res.*, 37(Web Server):W510–W514, Jul 2009a. ISSN 1362-4962. doi: 10.1093/nar/gkp322. URL <http://dx.doi.org/10.1093/nar/gkp322>.
- P. Benkert, M. Biasini, and T. Schwede. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*, 27(3):343–350, Feb 2011. ISSN 1460-2059. doi: 10.1093/bioinformatics/btq662. URL <http://dx.doi.org/10.1093/bioinformatics/btq662>.
- Pascal Benkert, Silvio C. E. Tosatto, and Dietmar Schomburg. Qmean: A comprehensive scoring function for model quality assessment. *Proteins: Struct., Funct., Bioinf.*, 71(1):261–277, Apr 2008. ISSN 1097-0134. doi: 10.1002/prot.21715. URL <http://dx.doi.org/10.1002/prot.21715>.
- Pascal Benkert, Torsten Schwede, and Silvio CE Tosatto. Qmeanclust: estimation of protein model quality by combining a composite scoring function with structural density information. *BMC Struct. Biol.*, 9(1):35, 2009b. ISSN 1472-6807. doi: 10.1186/1472-6807-9-35. URL <http://dx.doi.org/10.1186/1472-6807-9-35>.
- M. Berjanskii, P. Tang, J. Liang, J. A. Cruz, J. Zhou, Y. Zhou, E. Bassett, C. MacDonell, P. Lu, G. Lin, and et al. Genmr: a web server for rapid nmr-based protein structure

- determination. *Nucleic Acids Res.*, 37(Web Server):W670–W677, Jul 2009. ISSN 1362-4962. doi: 10.1093/nar/gkp280. URL <http://dx.doi.org/10.1093/nar/gkp280>.
- M. Berjanskii, Y. Liang, J. Zhou, P. Tang, P. Stothard, Y. Zhou, J. Cruz, C. MacDonell, G. Lin, P. Lu, and et al. Prossess: a protein structure evaluation suite and server. *Nucleic Acids Res.*, 38(Web Server):W633–W640, Jul 2010. ISSN 1362-4962. doi: 10.1093/nar/gkq375. URL <http://dx.doi.org/10.1093/nar/gkq375>.
- Frances C. Bernstein, Thomas F. Koetzle, Grahame J. B. Williams, Edgar F. Meyer, Michael D. Brice, John R. Rodgers, Olga Kennard, Takehiko Shimanouchi, and Mitsuo Tasumi. The protein data bank. a computer-based archival file for macromolecular structures. *Eur. J. Biochem.*, 80(2):319–324, Nov 1977. ISSN 1432-1033. doi: 10.1111/j.1432-1033.1977.tb11885.x. URL <http://dx.doi.org/10.1111/j.1432-1033.1977.tb11885.x>.
- M. Biasini, S. Bienert, A. Waterhouse, K. Arnold, G. Studer, T. Schmidt, F. Kiefer, T. G. Cassarino, M. Bertoni, L. Bordoli, and et al. Swiss-model: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.*, Apr 2014. ISSN 1362-4962. doi: 10.1093/nar/gku340. URL <http://dx.doi.org/10.1093/nar/gku340>.
- Brendan Borrell. Fraud rocks protein community. *Nature*, 462(7276):970–970, Dec 2009. ISSN 1476-4687. doi: 10.1038/462970a. URL <http://dx.doi.org/10.1038/462970a>.
- J. Bowie, R Luthy, and D Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253(5016):164–170, Jul 1991. ISSN 1095-9203. doi: 10.1126/science.1853201. URL <http://dx.doi.org/10.1126/science.1853201>.
- Carl-Ivar Brändén and T. Alwyn Jones. Between objectivity and subjectivity. *Nature*, 343(6260):687–689, Feb 1990. ISSN 0028-0836. doi: 10.1038/343687a0. URL <http://dx.doi.org/10.1038/343687a0>.
- Axel T. Brünger. Free r value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, 355(6359):472–475, Jan 1992. ISSN 0028-0836. doi: 10.1038/355472a0. URL <http://dx.doi.org/10.1038/355472a0>.
- Axel T. Brunger. [19] free r value: Cross-validation in crystallography. *Macromolecular Crystallography Part B*, pages 366–396, 1997. ISSN 0076-6879. doi: 10.1016/S0076-6879(97)77021-6. URL [http://dx.doi.org/10.1016/S0076-6879\(97\)77021-6](http://dx.doi.org/10.1016/S0076-6879(97)77021-6).

- Robert Bryll, Ricardo Gutierrez-Osuna, and Francis Quek. Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. *Pattern Recognition*, 36(6):1291–1302, Jun 2003. ISSN 0031-3203. doi: 10.1016/s0031-3203(02)00121-8. URL [http://dx.doi.org/10.1016/S0031-3203\(02\)00121-8](http://dx.doi.org/10.1016/S0031-3203(02)00121-8).
- L. Chiche, L. M. Gregoret, F. E. Cohen, and P. A. Kollman. Protein model structure evaluation using the solvation free energy of folding. *Proceedings of the National Academy of Sciences*, 87(8):3240–3243, Apr 1990. ISSN 1091-6490. doi: 10.1073/pnas.87.8.3240. URL <http://dx.doi.org/10.1073/pnas.87.8.3240>.
- James J. Chou, Shipeng Li, Claude B. Klee, and Ad Bax. Solution structure of  $\text{Ca}^{2+}$ -calmodulin reveals flexible hand-like properties of its domains. *Nat. Struct. Biol.*, 8(11):990–997, Nov 2001. ISSN 1072-8368. doi: 10.1038/nsb1101-990. URL <http://dx.doi.org/10.1038/nsb1101-990>.
- Cecilia Clementi. Coarse-grained models of protein folding: toy models or predictive tools? *Curr. Opin. Struct. Biol.*, 18(1):10–15, Feb 2008. doi: 10.1016/j.sbi.2007.10.005. URL <http://dx.doi.org/10.1016/j.sbi.2007.10.005>.
- Domenico Cozzetto, Andriy Kryshchak, Michele Ceriani, and Anna Tramontano. Assessment of predictions in the model quality assessment category. *Proteins: Struct., Funct., Bioinf.*, 69(S8):175–183, 2007. ISSN 1097-0134. doi: 10.1002/prot.21669. URL <http://dx.doi.org/10.1002/prot.21669>.
- K. A. Dill and J. L. MacCallum. The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046, Nov 2012. ISSN 1095-9203. doi: 10.1126/science.1219021. URL <http://dx.doi.org/10.1126/science.1219021>.
- Francisco S. Domingues, Peter Lackner, Antonina Andreeva, and Manfred J. Sippl. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J. Mol. Biol.*, 297(4):1003–1013, Apr 2000. ISSN 0022-2836. doi: 10.1006/jmbi.2000.3615. URL <http://dx.doi.org/10.1006/jmbi.2000.3615>.
- J Doye and W Poon. Protein crystallization in vivo. *Current Opinion in Colloid & Interface Science*, 11(1):40–46, Apr 2006. ISSN 1359-0294. doi: 10.1016/j.cocis.2005.10.002. URL <http://dx.doi.org/10.1016/j.cocis.2005.10.002>.
- F. Dressel. *Sequenz, Energie, Struktur - Untersuchungen zur Beziehung zwischen Primär- und Tertiärstruktur in globulären und Membran-Proteinen*. PhD thesis, Dresden, 2008.
- A. Keith Dunker and Zoran Obradovic. The protein trinity - linking function and disorder. *Nat. Biotechnol.*, 19(9):805–806, Sep 2001. ISSN 1087-0156. doi: 10.1038/nbt0901-805. URL <http://dx.doi.org/10.1038/nbt0901-805>.

- Keith Dunker and Richard Kriwacki. The orderly chaos of proteins. *Sci. Amer.*, 4:30–35, 2011.
- D. Eisenberg, R. Lüthy, and J. U. Bowie. Verify3d: assessment of protein models with three-dimensional profiles. *Methods Enzymol.*, 277:396–404, 1997.
- R. A. Engh and R. Huber. Accurate bond and angle parameters for x-ray protein structure refinement. *Acta Cryst A*, 47(4):392–400, Jul 1991. ISSN 0108-7673. doi: 10.1107/s0108767391001071. URL <http://dx.doi.org/10.1107/S0108767391001071>.
- Timothy D. Fenn, Dagmar Ringe, and Gregory A. Petsko. Xylose isomerase in substrate and inhibitor michaelis states: Atomic resolution studies of a metal-mediated hydride shift. *Biochemistry (Mosc.)*, 43(21):6464–6474, Jun 2004. ISSN 1520-4995. doi: 10.1021/bi049812o. URL <http://dx.doi.org/10.1021/bi049812o>.
- Emil Fischer. Einfluss der configuration auf die wirkung der enzyme. *Ber. Dtsch. Chem. Ges.*, 27(3):2985–2993, Oct 1894. ISSN 1099-0682. doi: 10.1002/cber.18940270364. URL <http://dx.doi.org/10.1002/cber.18940270364>.
- Lucy R. Forrest, Christopher L. Tang, and Barry Honig. On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. *Biophys. J.*, 91(2):508–517, Jul 2006. ISSN 0006-3495. doi: 10.1529/biophysj.106.082313. URL <http://dx.doi.org/10.1529/biophysj.106.082313>.
- E. Frank, M. Hall, L. Trigg, G. Holmes, and I. H. Witten. Data mining in bioinformatics using weka. *Bioinformatics*, 20(15):2479–2481, Oct 2004. ISSN 1460-2059. doi: 10.1093/bioinformatics/bth261. URL <http://dx.doi.org/10.1093/bioinformatics/bth261>.
- Jean Garnier, Jean-Francois Gibrat, and Barry Robson. [32] gor method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.*, pages 540–553, 1996. ISSN 0076-6879. doi: 10.1016/s0076-6879(96)66034-0. URL [http://dx.doi.org/10.1016/S0076-6879\(96\)66034-0](http://dx.doi.org/10.1016/S0076-6879(96)66034-0).
- Patrick Gendron, Sebastien Lemieux, and Francois Major. Quantitative analysis of nucleic acid three-dimensional structures. *J. Mol. Biol.*, 308(5):919–936, May 2001. ISSN 0022-2836. doi: 10.1006/jmbi.2001.4626. URL <http://dx.doi.org/10.1006/jmbi.2001.4626>.
- J.-F. Gibrat, J. Garnier, and B. Robson. Further developments of protein secondary structure prediction using information theory. *J. Mol. Biol.*, 198(3):425–443, Dec 1987. ISSN 0022-2836. doi: 10.1016/0022-2836(87)90292-0. URL [http://dx.doi.org/10.1016/0022-2836\(87\)90292-0](http://dx.doi.org/10.1016/0022-2836(87)90292-0).

- A. Giorgetti, D. Raimondo, A. E. Miele, and A. Tramontano. Evaluating the usefulness of protein structure models for molecular replacement. *Bioinformatics*, 21(Suppl 2): ii72–ii76, Sep 2005. ISSN 1460-2059. doi: 10.1093/bioinformatics/bti1112. URL <http://dx.doi.org/10.1093/bioinformatics/bti1112>.
- Michael Gribskov, Roland Lothy, and David Eisenberg. [9] profile analysis. *Methods Enzymol.*, pages 146–159, 1990. ISSN 0076-6879. doi: 10.1016/0076-6879(90)83011-w. URL [http://dx.doi.org/10.1016/0076-6879\(90\)83011-w](http://dx.doi.org/10.1016/0076-6879(90)83011-w).
- Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software. *SIGKDD Explor. Newsl.*, 11(1):10, Nov 2009. ISSN 1931-0145. doi: 10.1145/1656274.1656278. URL <http://dx.doi.org/10.1145/1656274.1656278>.
- F. Heinke, S. Schildbach, D. Stockmann, and D. Labudde. epros—a database and toolbox for investigating protein sequence-structure-function relationships through energy profiles. *Nucleic Acids Res.*, 41(D1):D320–D326, Jan 2013. ISSN 1362-4962. doi: 10.1093/nar/gks1079. URL <http://dx.doi.org/10.1093/nar/gks1079>.
- Florian Heinke and Dirk Labudde. Membrane protein stability analyses by means of protein energy profiles in case of nephrogenic diabetes insipidus. *Computational and Mathematical Methods in Medicine*, 2012:1–11, 2012. ISSN 1748-6718. doi: 10.1155/2012/790281. URL <http://dx.doi.org/10.1155/2012/790281>.
- Manfred Hendlich, Peter Lackner, Sabine Weitckus, Hannes Floeckner, Rosina Froschauer, Karl Gottsbacher, Georg Casari, and Manfred J. Sippl. Identification of native protein folds amongst a large number of incorrect models. *J. Mol. Biol.*, 216(1): 167–180, Nov 1990. ISSN 0022-2836. doi: 10.1016/s0022-2836(05)80068-3. URL [http://dx.doi.org/10.1016/S0022-2836\(05\)80068-3](http://dx.doi.org/10.1016/S0022-2836(05)80068-3).
- Tin Kam Ho. The random subspace method for constructing decision forests. 20(8): 832–844, 1998. ISSN 0162-8828. doi: 10.1109/34.709601. URL <http://dx.doi.org/10.1109/34.709601>.
- R. C. G. Holland, T. A. Down, M. Pocock, A. Prlic, D. Huen, K. James, S. Foisy, A. Drager, A. Yates, M. Heuer, and et al. Biojava: an open-source framework for bioinformatics. *Bioinformatics*, 24(18):2096–2097, Sep 2008. ISSN 1460-2059. doi: 10.1093/bioinformatics/btn397. URL <http://dx.doi.org/10.1093/bioinformatics/btn397>.
- G. Holmes, A. Donkin, and I.H. Witten. Weka: a machine learning workbench. *Proceedings of ANZIIS 94 - Australian New Zealand Intelligent Information Systems Conference*, pages 357–361, 1994. doi: 10.1109/anziis.1994.396988. URL <http://dx.doi.org/10.1109/ANZIIS.1994.396988>.

- D. T. Jones, W. R. Taylor, and J. M. Thornton. A new approach to protein fold recognition. *Nature*, 358(6381):86–89, Jul 1992. ISSN 0028-0836. doi: 10.1038/358086a0. URL <http://dx.doi.org/10.1038/358086a0>.
- R. P. Joosten, T. A. H. te Beek, E. Krieger, M. L. Hekkelman, R. W. W. Hooft, R. Schneider, C. Sander, and G. Vriend. A series of pdb related databases for everyday needs. *Nucleic Acids Res.*, 39(Database):D411–D419, Jan 2011. ISSN 1362-4962. doi: 10.1093/nar/gkq1105. URL <http://dx.doi.org/10.1093/nar/gkq1105>.
- W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr Sect A Cryst Phys Diffr Theor Gen Crystallogr*, 32(5):922–923, Sep 1976. ISSN 0567-7394. doi: 10.1107/s0567739476001873. URL <http://dx.doi.org/10.1107/S0567739476001873>.
- W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr Sect A Cryst Phys Diffr Theor Gen Crystallogr*, 34(5):827–828, Sep 1978. ISSN 0567-7394. doi: 10.1107/s0567739478001680. URL <http://dx.doi.org/10.1107/S0567739478001680>.
- Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, Dec 1983. ISSN 1097-0282. doi: 10.1002/bip.360221211. URL <http://dx.doi.org/10.1002/bip.360221211>.
- Jie Kang, Hans-Georg Lemaire, Axel Unterbeck, J. Michael Salbaum, Colin L. Masters, Karl-Heinz Grzeschik, Gerd Multhaup, Konrad Beyreuther, and Benno Müller-Hill. The precursor of alzheimer's disease amyloid a4 protein resembles a cell-surface receptor. *Nature*, 325(6106):733–736, Feb 1987. ISSN 0028-0836. doi: 10.1038/325733a0. URL <http://dx.doi.org/10.1038/325733a0>.
- J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666, Mar 1958. ISSN 0028-0836. doi: 10.1038/181662a0. URL <http://dx.doi.org/10.1038/181662a0>.
- Daisuke Kihara, Hao Chen, and Yifeng Yang. Quality assessment of protein structure models. *Current Protein & Peptide Science*, 10(3):216–228, Jun 2009. ISSN 1389-2037. doi: 10.2174/138920309788452173. URL <http://dx.doi.org/10.2174/138920309788452173>.
- Yujin E. Kim, Mark S. Hipp, Andreas Bracher, Manajit Hayer-Hartl, and F. Ulrich Hartl. Molecular chaperone functions in protein folding and proteostasis. *Annu. Rev. Biochem.*, 82(1):323–355, Jun 2013. ISSN 1545-4509. doi:



- 10.1146/annurev-biochem-060208-092442. URL <http://dx.doi.org/10.1146/annurev-biochem-060208-092442>.
- Gerard J. Kleywegt and T. Alwyn Jones. Databases in protein crystallography. *Acta Cryst D*, 54(6):1119–1131, Nov 1998. ISSN 0907-4449. doi: 10.1107/s0907444998007100. URL <http://dx.doi.org/10.1107/S0907444998007100>.
- A. Kloczkowski, K.-L. Ting, R.L. Jernigan, and J. Garnier. Combining the gor v algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. *Proteins*, 49(2):154–166, Nov 2002. ISSN 1097-0134. doi: 10.1002/prot.10181. URL <http://dx.doi.org/10.1002/prot.10181>.
- J. J. Kovacevic. Computational analysis of position-dependent disorder content in disprot database. *Genomics, Proteomics & Bioinformatics*, 10(3):158–165, Jun 2012. ISSN 1672-0229. doi: 10.1016/j.gpb.2012.01.002. URL <http://dx.doi.org/10.1016/j.gpb.2012.01.002>.
- Nadezda V. Kovalevskaya, Michiel Waterbeemd, Fedir M. Bokhovchuk, Neil Bate, Rene J. M. Bindels, Joost G. J. Hoenderop, and Geerten W. Vuister. Structural analysis of calmodulin binding to ion channels demonstrates the role of its plasticity in regulation. *Pfluegers Archiv - European Journal of Physiology*, 465(11):1507–1519, Apr 2013. ISSN 1432-2013. doi: 10.1007/s00424-013-1278-0. URL <http://dx.doi.org/10.1007/s00424-013-1278-0>.
- E. Krissinel and K. Henrick. Secondary-structure matching (ssm), a new tool for fast protein structure alignment in three dimensions. *Acta Cryst D*, 60(12):2256–2268, Nov 2004. ISSN 0907-4449. doi: 10.1107/s0907444904026460. URL <http://dx.doi.org/10.1107/S0907444904026460>.
- Andriy Kryshtafovych and Krzysztof Fidelis. Protein structure prediction and model quality assessment. *Drug Discovery Today*, 14(7-8):386–393, Apr 2009. ISSN 1359-6446. doi: 10.1016/j.drudis.2008.11.010. URL <http://dx.doi.org/10.1016/j.drudis.2008.11.010>.
- Andriy Kryshtafovych, Krzysztof Fidelis, and Anna Tramontano. Evaluation of model quality predictions in CASP9. *Proteins: Struct., Funct., Bioinf.*, 79(S10):91–106, 2011. ISSN 0887-3585. doi: 10.1002/prot.23180. URL <http://dx.doi.org/10.1002/prot.23180>.
- Andriy Kryshtafovych, Alessandro Barbato, Krzysztof Fidelis, Bohdan Monastyrskyy, Torsten Schwede, and Anna Tramontano. Assessment of the assessment: Evaluation of the model quality estimates in CASP10. *Proteins: Struct., Funct., Bioinf.*, 82:112–126, Feb 2014a. ISSN 0887-3585. doi: 10.1002/prot.24347. URL <http://dx.doi.org/10.1002/prot.24347>.

- Andriy Kryshtafovych, Bohdan Monastyrskyy, and Krzysztof Fidelis. CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. *Proteins: Struct., Funct., Bioinf.*, 82:7–13, Feb 2014b. ISSN 0887-3585. doi: 10.1002/prot.24399. URL <http://dx.doi.org/10.1002/prot.24399>.
- I. D. Kuntz. Structure-based strategies for drug design and discovery. *Science*, 257(5073):1078–1082, Aug 1992. ISSN 1095-9203. doi: 10.1126/science.257.5073.1078. URL <http://dx.doi.org/10.1126/science.257.5073.1078>.
- R. A. Laskowski. Pdbsum: summaries and analyses of pdb structures. *Nucleic Acids Res.*, 29(1):221–222, Jan 2001. ISSN 1362-4962. doi: 10.1093/nar/29.1.221. URL <http://dx.doi.org/10.1093/nar/29.1.221>.
- R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton. Procheck: a program to check the stereochemical quality of protein structures. *J Appl Cryst*, 26(2):283–291, Apr 1993. ISSN 0021-8898. doi: 10.1107/s0021889892009944. URL <http://dx.doi.org/10.1107/S0021889892009944>.
- Roman A. Laskowski. Structural quality assurance. *Methods Biochem. Anal.*, pages 273–303, Jan 2005. ISSN 1934-4325. doi: 10.1002/0471721204.ch14. URL <http://dx.doi.org/10.1002/0471721204.ch14>.
- Roman A. Laskowski, J. Antoon C. Rullmann, Malcolm W. MacArthur, Robert Kaptein, and Janet M. Thornton. Aqua and procheck-nmr: Programs for checking the quality of protein structures solved by nmr. *J. Biomol. NMR*, 8(4), Dec 1996. ISSN 1573-5001. doi: 10.1007/bf00228148. URL <http://dx.doi.org/10.1007/BF00228148>.
- Yaohang Li. Conformational sampling in template-free protein loop structure modeling: An overview. *Computational and Structural Biotechnology Journal*, 5(6), Feb 2013. ISSN 2001-0370. doi: 10.5936/csbj.201302003. URL <http://dx.doi.org/10.5936/csbj.201302003>.
- Tianyun Liu, Grace W. Tang, and Emidio Capriotti. Comparative modeling: The state of the art and protein drug target structure prediction. *CCHTS*, 14(6):532–547, Jul 2011. ISSN 1386-2073. doi: 10.2174/138620711795767811. URL <http://dx.doi.org/10.2174/138620711795767811>.
- R. Maiti, G. H. Van Domselaar, H. Zhang, and D. S. Wishart. Superpose: a simple server for sophisticated structural superposition. *Nucleic Acids Res.*, 32(Web Server):W590–W594, Jul 2004. ISSN 1362-4962. doi: 10.1093/nar/gkh477. URL <http://dx.doi.org/10.1093/nar/gkh477>.
- L. J. McGuffin, K. Bryson, and D. T. Jones. The psipred protein structure prediction server. *Bioinformatics*, 16(4):404–405, Apr 2000. ISSN 1460-2059. doi: 10.1093/

- bioinformatics/16.4.404. URL <http://dx.doi.org/10.1093/bioinformatics/16.4.404>.
- F. Melo, D. Devos, E. Depiereux, and E. Feytmans. Anolea: A www server to assess protein structures. *Proc Int Conf Intell Syst Mol Biol.*, 5:187–90, 1997.
- Francisco Melo and Ernest Feytmans. Novel knowledge-based mean force potential at atomic level. *J. Mol. Biol.*, 267(1):207–222, Mar 1997. ISSN 0022-2836. doi: 10.1006/jmbi.1996.0868. URL <http://dx.doi.org/10.1006/jmbi.1996.0868>.
- Francisco Melo and Ernest Feytmans. Assessing protein structures with a non-local atomic interaction energy. *J. Mol. Biol.*, 277(5):1141–1152, Apr 1998. ISSN 0022-2836. doi: 10.1006/jmbi.1998.1665. URL <http://dx.doi.org/10.1006/jmbi.1998.1665>.
- Ethan A. Merritt. Comparing anisotropic displacement parameters in protein structures. *Acta Cryst D*, 55(12):1997–2004, Dec 1999. ISSN 0907-4449. doi: 10.1107/S0907444999011853. URL <http://dx.doi.org/10.1107/S0907444999011853>.
- Susan Miller, Joel Janin, Arthur M. Lesk, and Cyrus Chothia. Interior and surface of monomeric proteins. *J. Mol. Biol.*, 196(3):641–656, Aug 1987. ISSN 0022-2836. doi: 10.1016/0022-2836(87)90038-6. URL [http://dx.doi.org/10.1016/0022-2836\(87\)90038-6](http://dx.doi.org/10.1016/0022-2836(87)90038-6).
- A. E. Mirsky and Linus Pauling. On the structure of native, denatured, and coagulated proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 22(7):439–447, 1936.
- Anne Louise Morris, Malcolm W. MacArthur, E. Gail Hutchinson, and Janet M. Thornton. Stereochemical quality of protein structure coordinates. *Proteins*, 12(4):345–364, Apr 1992. ISSN 1097-0134. doi: 10.1002/prot.340120407. URL <http://dx.doi.org/10.1002/prot.340120407>.
- John Moult, Jan T. Pedersen, Richard Judson, and Krzysztof Fidelis. A large-scale experiment to assess protein structure prediction methods. *Proteins*, 23(3):ii–iv, Nov 1995. ISSN 1097-0134. doi: 10.1002/prot.340230303. URL <http://dx.doi.org/10.1002/prot.340230303>.
- John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, and Anna Tramontano. Critical assessment of methods of protein structure prediction (CASP)-round ix. *Proteins: Struct., Funct., Bioinf.*, 79(S10):1–5, Oct 2011. doi: 10.1002/prot.23200. URL <http://dx.doi.org/10.1002/prot.23200>.
- John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Anna Tramontano. Critical assessment of methods of protein structure prediction (CASP) -

- round x. *Proteins: Struct., Funct., Bioinf.*, 82:1–6, Feb 2014. ISSN 0887-3585. doi: 10.1002/prot.24452. URL <http://dx.doi.org/10.1002/prot.24452>.
- Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48(3):443–453, Mar 1970. ISSN 0022-2836. doi: 10.1016/0022-2836(70)90057-4. URL [http://dx.doi.org/10.1016/0022-2836\(70\)90057-4](http://dx.doi.org/10.1016/0022-2836(70)90057-4).
- David Nelson and Michael Cox. *Lehninger Biochemie (Springer-Lehrbuch) (German Edition)*. Springer, 2010. ISBN 3540686371. URL <http://www.amazon.com/Lehninger-Biochemie-Springer-Lehrbuch-German-Edition/dp/3540686371%3FSubscriptionId%3D0JYN1NVW651KCA56C102%26tag%3Dtechkie-20%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D3540686371>.
- T. Noguchi. Pdb-reprdb: a database of representative protein chains from the protein data bank (pdb). *Nucleic Acids Res.*, 29(1):219–220, Jan 2001. ISSN 1362-4962. doi: 10.1093/nar/29.1.219. URL <http://dx.doi.org/10.1093/nar/29.1.219>.
- Chris Oostenbrink, Alessandra Villa, Alan E. Mark, and Wilfred F. Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The gromos force-field parameter sets 53a5 and 53a6. *J. Comput. Chem.*, 25(13):1656–1676, Oct 2004. ISSN 1096-987X. doi: 10.1002/jcc.20090. URL <http://dx.doi.org/10.1002/jcc.20090>.
- Pance Panov and Saso Dzeroski. Combining bagging and random subspaces to create better ensembles. *Lecture Notes in Computer Science*, pages 118–129, 2007. ISSN 1611-3349. doi: 10.1007/978-3-540-74825-0\_11. URL [http://dx.doi.org/10.1007/978-3-540-74825-0\\_11](http://dx.doi.org/10.1007/978-3-540-74825-0_11).
- William R. Pearson. [5] rapid and sensitive sequence comparison with fastp and fasta. *Methods Enzymol.*, pages 63–98, 1990. ISSN 0076-6879. doi: 10.1016/0076-6879(90)83007-v. URL [http://dx.doi.org/10.1016/0076-6879\(90\)83007-v](http://dx.doi.org/10.1016/0076-6879(90)83007-v).
- Gianluca Pollastri, Pierre Baldi, Pietro Fariselli, and Rita Casadio. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins*, 47(2):142–153, May 2002. ISSN 1097-0134. doi: 10.1002/prot.10069. URL <http://dx.doi.org/10.1002/prot.10069>.
- A. Prlic, A. Yates, S. E. Bliven, P. W. Rose, J. Jacobsen, P. V. Troshin, M. Chapman, J. Gao, C. H. Koh, S. Foisy, and et al. Biojava: an open-source framework for bioinformatics in 2012. *Bioinformatics*, 28(20):2693–2695, Oct 2012. ISSN 1460-

2059. doi: 10.1093/bioinformatics/bts494. URL <http://dx.doi.org/10.1093/bioinformatics/bts494>.

Domenico Raimondo, Alejandro Giorgetti, Alejandro Giorgetti, Stefania Bosi, and Anna Tramontano. Automatic procedure for using models of proteins in molecular replacement. *Proteins: Struct., Funct., Bioinf.*, 66(3):689–696, Nov 2006. ISSN 0887-3585. doi: 10.1002/prot.21225. URL <http://dx.doi.org/10.1002/prot.21225>.

G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, 7(1):95–99, Jul 1963. ISSN 0022-2836. doi: 10.1016/s0022-2836(63)80023-6. URL [http://dx.doi.org/10.1016/S0022-2836\(63\)80023-6](http://dx.doi.org/10.1016/S0022-2836(63)80023-6).

Randy J. Read and Gayatri Chavali. Assessment of CASP7 predictions in the high accuracy template-based modeling category. *Proteins: Struct., Funct., Bioinf.*, 69(S8):27–37, 2007. ISSN 1097-0134. doi: 10.1002/prot.21662. URL <http://dx.doi.org/10.1002/prot.21662>.

Randy J. Read, Paul D. Adams, W. Bryan Arendall, Axel T. Brunger, Paul Emsley, Robbie P. Joosten, Gerard J. Kleywegt, Eugene B. Krissinel, Thomas Lutteke, Zbyszek Otwinowski, and et al. A new generation of crystallographic validation tools for the protein data bank. *Structure*, 19(10):1395–1412, Oct 2011. ISSN 0969-2126. doi: 10.1016/j.str.2011.08.006. URL <http://dx.doi.org/10.1016/j.str.2011.08.006>.

Gale Rhodes. *Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models Third Edition*. Elsevier, 2006.

F M Richards. Areas, volumes, packing, and protein structure. *Annu. Rev. Biophys. Bioeng.*, 6(1):151–176, Jun 1977. ISSN 0084-6589. doi: 10.1146/annurev.bb.06.060177.001055. URL <http://dx.doi.org/10.1146/annurev.bb.06.060177.001055>.

M. I. Sadowski and D. T. Jones. Benchmarking template selection and model quality assessment for high-resolution comparative modeling. *Proteins: Struct., Funct., Bioinf.*, 69(3):476–485, Jul 2007. ISSN 0887-3585. doi: 10.1002/prot.21531. URL <http://dx.doi.org/10.1002/prot.21531>.

Manfred J. Sippl. Calculation of conformational ensembles from potentials of mean force. *J. Mol. Biol.*, 213(4):859–883, Jun 1990. ISSN 0022-2836. doi: 10.1016/s0022-2836(05)80269-4. URL [http://dx.doi.org/10.1016/S0022-2836\(05\)80269-4](http://dx.doi.org/10.1016/S0022-2836(05)80269-4).

Manfred J. Sippl. Boltzmann’s principle, knowledge-based mean fields and protein folding. an approach to the computational determination of protein structures.

- J. Computer-Aided Mol. Des.*, 7(4):473–501, Aug 1993a. ISSN 1573-4951. doi: 10.1007/bf02337562. URL <http://dx.doi.org/10.1007/BF02337562>.
- Manfred J. Sippl. Recognition of errors in three-dimensional structures of proteins. *Proteins*, 17(4):355–362, Dec 1993b. ISSN 1097-0134. doi: 10.1002/prot.340170404. URL <http://dx.doi.org/10.1002/prot.340170404>.
- Manfred J Sippl. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.*, 5(2):229–235, Apr 1995. ISSN 0959-440X. doi: 10.1016/0959-440x(95)80081-6. URL [http://dx.doi.org/10.1016/0959-440X\(95\)80081-6](http://dx.doi.org/10.1016/0959-440X(95)80081-6).
- T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147(1):195–197, Mar 1981. ISSN 0022-2836. doi: 10.1016/0022-2836(81)90087-5. URL [http://dx.doi.org/10.1016/0022-2836\(81\)90087-5](http://dx.doi.org/10.1016/0022-2836(81)90087-5).
- J. Soding. Protein homology detection by HMM-hmm comparison. *Bioinformatics*, 21(7):951–960, Nov 2004. ISSN 1460-2059. doi: 10.1093/bioinformatics/bti125. URL <http://dx.doi.org/10.1093/bioinformatics/bti125>.
- Janet Thornton. Structural genomics takes off. *Trends Biochem. Sci.*, 26(2):88–89, Feb 2001. ISSN 0968-0004. doi: 10.1016/s0968-0004(00)01765-5. URL [http://dx.doi.org/10.1016/S0968-0004\(00\)01765-5](http://dx.doi.org/10.1016/S0968-0004(00)01765-5).
- V. N. Uversky. Natively unfolded proteins: A point where biology waits for physics. *Protein Sci.*, 11(4):739–756, Apr 2002. ISSN 1469-896X. doi: 10.1110/ps.4210102. URL <http://dx.doi.org/10.1110/ps.4210102>.
- G. Verkhivker, K. Appelt, S.T. Freer, and J.E. Villafranca. Empirical free energy calculations of ligand-protein crystallographic complexes. i. knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. *Protein Engineering Design and Selection*, 8(7):677–691, Jul 1995. ISSN 1741-0134. doi: 10.1093/protein/8.7.677. URL <http://dx.doi.org/10.1093/protein/8.7.677>.
- M. Wiederstein and M. J. Sippl. Prosa-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.*, 35(Web Server):W407–W410, May 2007. ISSN 1362-4962. doi: 10.1093/nar/gkm290. URL <http://dx.doi.org/10.1093/nar/gkm290>.
- L. Willard. Vadar: a web server for quantitative evaluation of protein structure quality. *Nucleic Acids Res.*, 31(13):3316–3319, Jul 2003. ISSN 1362-4962. doi: 10.1093/nar/gkg565. URL <http://dx.doi.org/10.1093/nar/gkg565>.
- Mark A. Wilson and Axel T. Brunger. The 1.0 Å crystal structure of ca<sup>2+</sup>-bound calmod-

ulin: an analysis of disorder and implications for functionally relevant plasticity. *J. Mol. Biol.*, 301(5):1237–1256, Sep 2000. ISSN 0022-2836. doi: 10.1006/jmbi.2000.4029. URL <http://dx.doi.org/10.1006/jmbi.2000.4029>.

Alexander Wlodawer, Wladek Minor, Zbigniew Dauter, and Mariusz Jaskolski. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS J.*, 275(1):1–21, Nov 2007. ISSN 1742-464X. doi: 10.1111/j.1742-4658.2007.06178.x. URL <http://dx.doi.org/10.1111/j.1742-4658.2007.06178.x>.

Kurt Wüthrich. Protein structure determination in solution by nmr spectroscopy. *J. Biol. Chem.*, 265(36):22059–22062, Dec 1990.

Hsien Wu. Studies on denaturation of proteins. xiii. a theory of denaturation. *Chin. J. Physiol.*, 5:321–344, 1931.

A. Zemla. Lga: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.*, 31(13):3370–3374, Jul 2003. ISSN 1362-4962. doi: 10.1093/nar/gkg571. URL <http://dx.doi.org/10.1093/nar/gkg571>.

Nan Zhao, Jing Ginger Han, Chi-Ren Shyu, and Dmitry Korkin. Determining effects of non-synonymous snps on protein-protein interactions using supervised and semi-supervised learning. *PLoS Comput. Biol.*, 10(5):e1003592, May 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003592. URL <http://dx.doi.org/10.1371/journal.pcbi.1003592>.





## Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht.

Diese Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt.

Mittweida, August 25, 2014